Unit 11

Regression

Introduction

So far in this module, models for variation have been developed that are appropriate for studying a suitably defined underlying population in its entirety. For instance, in Unit 6, you met an example of data collected on the heights of elderly women. (These data formed part of a study into the disease osteoporosis.) There was evident variation in the sample, and it was suggested that the variation in the data could be modelled adequately by a normal distribution with appropriately chosen values for its parameters μ and σ. This model provided useful information about the population of heights of all elderly women. However, it did not provide information about the population of heights of females in general – for example, the variation in heights of teenage girls is likely to be different from that for elderly women. In the wider population of women in general, we would expect height to depend on age at least up to about 15 or 16 years old; manufacturers of children's clothes, for instance, need to be aware of this relationship. The following example illustrates how such a relationship can be modelled.

Example 1 *Heights of schoolboys*

A very early study conducted for the Massachusetts Board of Health in 1877 recorded the age and height of each of 24 500 Boston schoolboys between the ages of 6 and 10 years. A histogram of the heights of the boys (in inches) is shown in Figure 1.

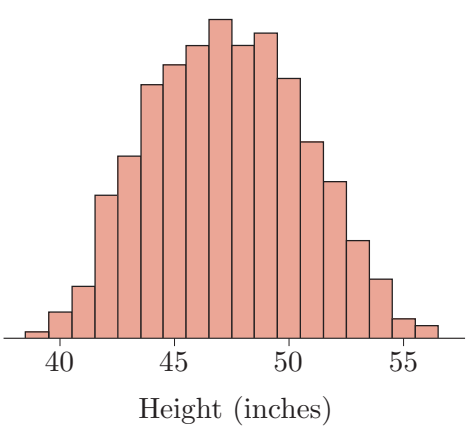


Figure 1 The variation in height for boys between the ages of 6 and 10 years

Nineteenth-century schoolboys (and girls)

As for the heights of elderly women, we could look for a model for the variation in the heights of these schoolboys; a normal distribution again seems a reasonable possibility, but with different values for its parameters μ and σ. This model would provide some general information about the heights of nineteenth-century Boston schoolboys between the ages of 6 and 10 years, but it would not tell us anything about the relationship between height and age for these boys. Instead, if the boys are divided into five age

Figures 1 and 2 are adapted from Peters, W.S. (1987) *Counting for Something – Statistical Principles and Personalities*, New York, Springer-Verlag, p. 90.

groups of a year each (ages 6 to 10 years) and a histogram is drawn separately for each group, then the same data may be represented as in Figure 2. In the figure, height is represented by the vertical axis while age, grouped into 6, 7, 8, 9 or 10 years, is represented by the horizontal axis. Each histogram is plotted on its side rather than in the usual way.

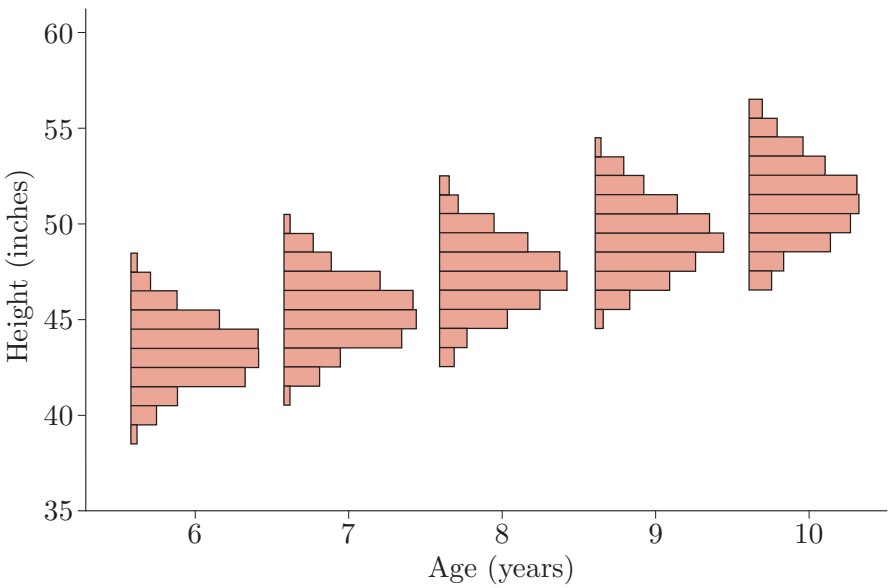


Figure 2 The variation in height for boys in each of the age groups 6 to 10 years

You can now see that, *for each age group*, a normal distribution might provide a good model for the variation in heights. It also appears that the mean height increases roughly linearly with age – at least between the ages of 6 and 10 years. (The increasing linear trend would not continue into all higher ages, of course.) The spread of heights about the mean does not seem to alter much with age, that is, the variance in heights appears to be approximately constant. So perhaps the variation in heights of nineteenth-century Boston schoolboys can be adequately modelled by a collection of normal distributions where the normal distributions differ with respect to their means, μ, but all have the same variance, σ^2. Moreover, rather than being an arbitrary collection of means, it seems that the means of the normal distributions increase linearly with age.

Relationships of this sort between variables are the subject of this unit. As you may know already, statistical models that reflect the way in which variation in an observed variable changes with one or more other variables are called **regression models**. The development and use of these models is known as **regression**. Situations where a regression model might be useful include the following.

- Economists predict future employment rates on the basis of past and current rates, together with various economic variables.
- Farmers wish to know how the yield of crops depends on the amount of fertiliser used.

- Doctors must decide, on the basis of particular measurements, how much of a drug to give to a patient.
- A car owner might be interested in knowing how driving her car at different speeds alters its fuel consumption.

In Section 1, a few more examples are given before the *general regression model* is formally defined. Also, a particularly important regression model, the *linear regression model*, is introduced. In Section 2, a method for fitting linear regression models to data is described. Section 3 is principally concerned with checking the modelling assumptions; at its end, you will see how to fit the linear regression model and how to check the assumptions using Minitab. Statistical inference, such as testing hypotheses and calculating confidence intervals, for linear regression models is discussed quite briefly in Section 4. The unit ends by introducing *multiple regression*, where the relationship between a variable and more than one other variable is of interest.

1 Regression models

1.1 Examples

This section begins with a few more examples of contexts in which regression data arise.

Cruise control

Example 2 *Driving at constant speed*

Consider the following hypothetical situation. For a car driving at a constant speed of 50 mph, the relationship between the distance travelled and the time spent driving can be represented by the straight line in Figure 3(a).

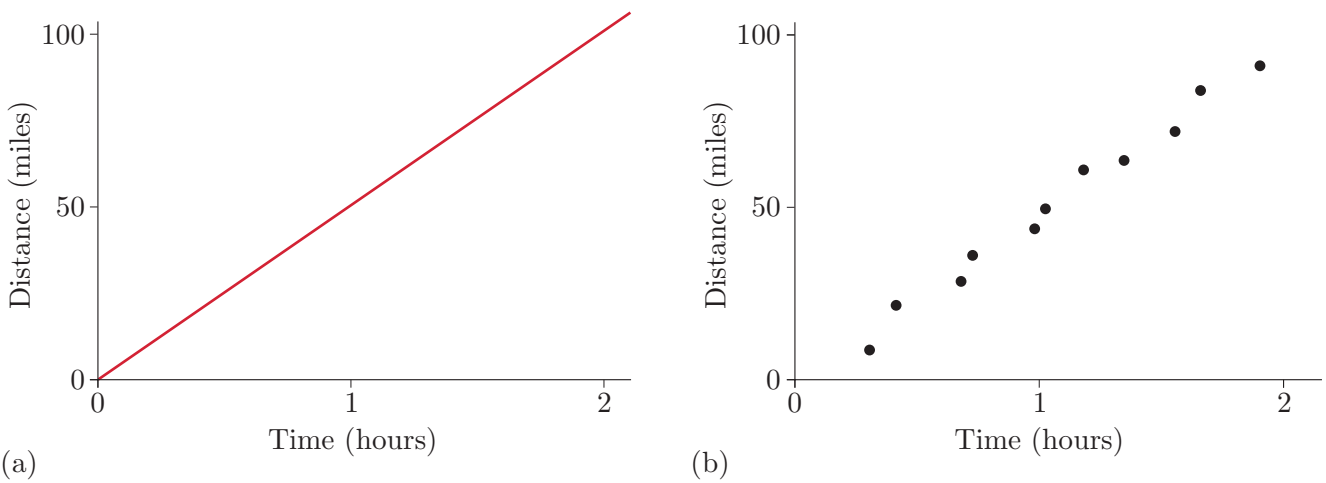


Figure 3 (a) Distance against time. (b) A scatterplot of 'real' observations.

A scatterplot of observations measured without error would consist of dots all lying exactly on the straight line in Figure 3(a). However, in a scatterplot of real observations, the dots are very unlikely to lie exactly along the straight line but would be scattered around the line, perhaps looking something like Figure 3(b). So we need a model that will describe the linear relationship underlying these data while at the same time allowing for some deviation of the data from the line.

Example 3 *Forbes's data on the boiling point of water*

Timby's Mercury Barometer, patented 1857

In the 1840s and 1850s the Scottish physicist James Forbes was interested in developing a method for estimating altitude on a hillside from measurement of the boiling point of water there. The temperature at which water boils is affected by atmospheric pressure which, in turn, is affected by altitude. (You might know that the higher the altitude, the lower the pressure, and the lower the boiling point of water.)

So boiling point depends on atmospheric pressure, and if the details of that relationship were known, Forbes concluded that it should be possible to turn the relationship round so that climbers could estimate their height from the temperature at which water boiled. Carrying barometers – which, at that time, were large instruments which included a long, thin glass tube containing mercury – up and down hills intact was a tricky business; boiling a pan of water and measuring the temperature of the boiling point was less troublesome. Here, however, we will concentrate on the initial question of the way boiling point depends on atmospheric pressure.

The data in Table 1 give the boiling point (in °F) and atmospheric pressure (in inches Hg – that is, inches of mercury) for 17 locations in the Alps and in Scotland.

Table 1 Forbes's data

Boiling point (°F)	194.5	194.3	197.9	198.4	199.4	199.9
Pressure (inches Hg)	20.79	20.79	22.40	22.67	23.15	23.35
Boiling point (°F)	200.9	201.1	201.4	201.3	203.6	204.6
Pressure (inches Hg)	23.89	23.99	24.02	24.01	25.14	26.57
Boiling point (°F)	209.5	208.6	210.7	211.9	212.2	
Pressure (inches Hg)	28.49	27.76	29.04	29.88	30.06	

(Source: Forbes, J.D. (1857) 'Further experiments and remarks on the measurement of heights by the boiling point of water', *Transactions of the Royal Society of Edinburgh*, vol. 21, no. 2, pp. 235–43)

The scatterplot of these data in Figure 4 suggests that there may well be a straight-line relationship between the boiling point of water and atmospheric pressure. A model for the data should exhibit this linear relationship.

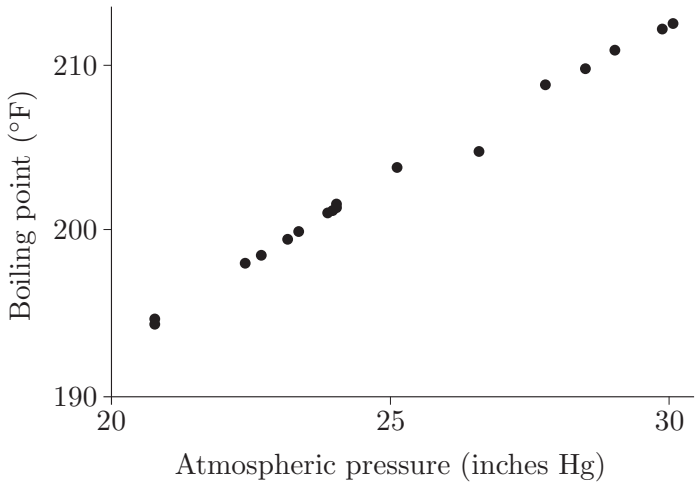


Figure 4 Boiling point of water against atmospheric pressure

Example 4 *The strength of timber beams*

A dataset contains the results of an investigation into how specific gravity (a measure of density) and moisture content might influence the strength of timber beams. Table 2 contains measurements of the three variables for each of ten beams. Unfortunately, units of measurement are not given in the source.

Table 2 Strength of beams

Strength	11.14	12.74	13.13	11.51	12.38
Specific gravity	0.499	0.558	0.604	0.441	0.550
Moisture content	11.1	8.9	8.8	8.9	8.8
Strength	12.60	11.13	11.70	11.02	11.42
Specific gravity	0.528	0.418	0.480	0.406	0.467
Moisture content	9.9	10.7	10.5	10.5	10.7

(Source: Draper, N.R. and Stoneman, D.M. (1966) 'Testing for the inclusion of variables in linear regression by a randomisation technique', *Technometrics*, vol. 8, no. 4, pp. 695–9)

The scatterplot of strength against specific gravity in Figure 5(a) (overleaf) suggests some sort of increasing linear relationship between strength and specific gravity, though possibly there is an outlier at $(0.499, 11.14)$.

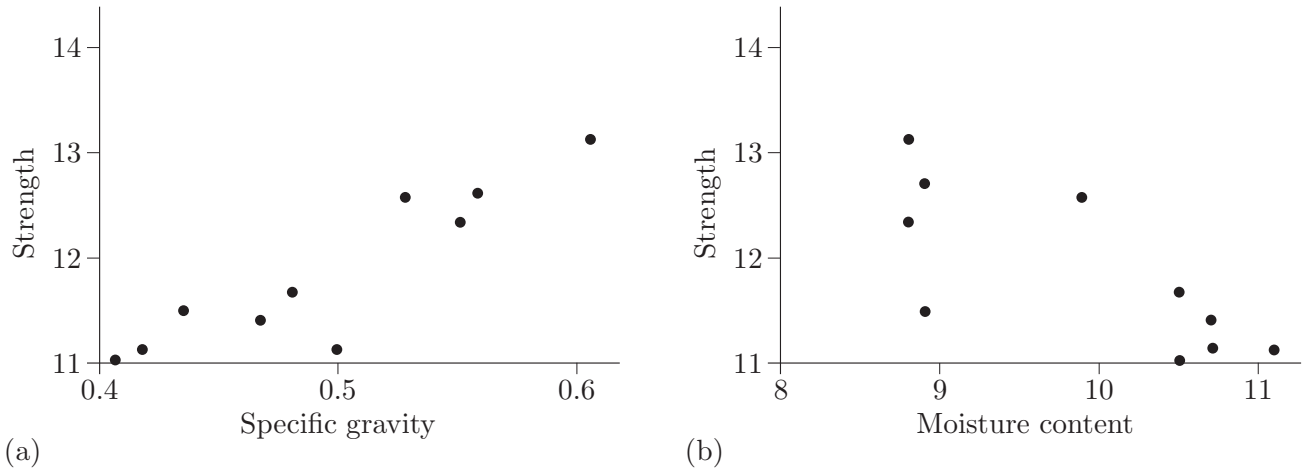
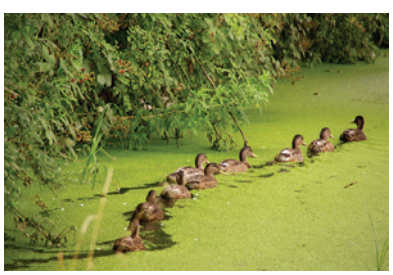


Figure 5 Scatterplots: (a) strength against specific gravity (b) strength against moisture content

Although the scatterplot of strength against moisture content in Figure 5(b) suggests an overall downward trend, it is not all that convincing – it does not seem to strongly suggest any particular form for a relationship. Linearity might, however, be as good as any.

Example 5 *The growth of duckweed*



Ducks in duckweed

In his 1917 book *On Growth and Form*, the Scottish mathematical biologist Sir D'Arcy Wentworth Thompson recounts an experiment into the growth of duckweed, a plant that grows on water. Growth was monitored by counting duckweed fronds at weekly intervals for eight weeks, starting one week after the introduction of a single duckweed plantlet into a growth medium (in this case, pure water). Initially (week 0) there were 20 fronds on the plantlet. The data are given in Table 3.

Table 3 Duckweed growth

Week	1	2	3	4	5	6	7	8
Fronds	30	52	77	135	211	326	550	1052

(Source: Thompson refers to work summarised in Bottomley, W.B. (1914) 'Some accessory factors in plant growth and nutrition', *Proceedings of the Royal Society, Series B*, vol. 88, no. 602, 237–47)

A scatterplot of the data is given in Figure 6.

You can see that there is a very strong suggestion of a relationship between duckweed growth and passing time; but, unlike in the previous examples, the relationship is not linear. Instead, it might be possible to fit a curve to the data – perhaps some sort of power or exponential function.

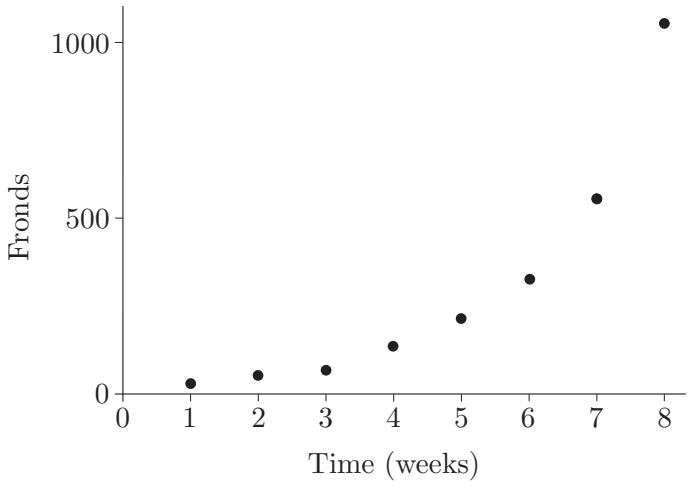


Figure 6 Duckweed growth

Often, data arise as the result of an experiment specifically designed to investigate the effect that changes in one variable (or more than one variable) have on another variable: Forbes investigated the effect of changes in atmospheric pressure on the boiling point of water (see Example 3); the strength of timber was measured for different values of specific gravity and moisture content (see Example 4); the scatterplot in Figure 6 suggests how the growth of duckweed depends on passing time. In all three examples, we could naturally think of one variable (or two variables, in Example 4) having an effect on, or '*explaining*', another variable. And in each of these three examples, it would not be at all natural or even sensible to swap the variables around: we would not speak of an increase or decrease in the boiling point of water 'changing' the atmospheric pressure; or of the strength of a timber beam 'having an effect on' the moisture content; or of the growth of duckweed 'causing' a change in time.

The word '*explain*' should not be taken too literally in this context; it is used only to express that a change in one variable has an effect on another variable.

A variable that 'explains' another variable is called an **explanatory variable**. In Example 3, Forbes used atmospheric pressure to 'explain' the boiling points of water at various altitudes. So atmospheric pressure is an explanatory variable. In Example 4, there are two explanatory variables, namely specific gravity and moisture content, both 'explaining' the strength of the timber beams. In Example 5, time 'explains' the changing number of duckweed fronds, so time is an explanatory variable.

In Example 3, the measured boiling point can be regarded as a *response* to a given atmospheric pressure. For different pressures, the boiling point will be different. The variable that 'responds' to the value of the explanatory variable is called the **response variable**. In Example 4, the response variable is the strength of the timber beam; and in Example 5, the response variable is the number of duckweed fronds.

Other names are sometimes used for the response and explanatory variables. These include *dependent variable* and *independent variable*, respectively. The explanatory variable is also called the *predictor variable*, the *regressor* or the *covariate*.

It can be argued that the name 'independent variable' is misleading because we are interested in relationships between variables which are not independent.

As you saw in Example 4, there can be more than one explanatory variable. However, first we will be concerned with the case where there is only one explanatory variable. In this case, the model is often referred to as a *simple* regression model. The word 'simple' here refers purely to the number of variables involved in the regression; you can be the judge of whether or not the interpretation, properties and application of the simple regression model are what you would call simple! When there are two or more explanatory variables, the model is called a *multiple* regression model. This will be the topic of Section 5.

Activity 1 *Heights of Boston schoolboys*

In the Introduction, data on the age and height of schoolboys from Boston were discussed. Which of the two variables, age and height, would you regard as the response variable and which as the explanatory variable?

Activity 2 *Paper strength*

Table 4 contains data on the strength of kraft paper. ('Kraft' refers to a method of paper production. The paper is of a thick brown type used for wrapping.) The tensile strength (in pounds per square inch (p.s.i.)) of the paper was measured along with the percentage of hardwood in the batch of pulp from which the paper was produced. In Figure 7, tensile strength is plotted against hardwood content.

Table 4 Paper strength

Strength (p.s.i.)	Hardwood content (%)
6.3	1.0
11.1	1.5
20.0	2.0
24.0	3.0
26.1	4.0
30.0	4.5
33.8	5.0
34.0	5.5
38.1	6.0
39.9	6.5
42.0	7.0
46.1	8.0
53.1	9.0
52.0	10.0
52.5	11.0
48.0	12.0
42.8	13.0
27.8	14.0
21.9	15.0

(Source: Joglekar, G., Schuenemeyer, J.H. and LaRiccia, V. (1989) 'Lack-of-fit testing when replicates are not available', *American Statistician*, vol. 43, no. 3, pp. 135–43)

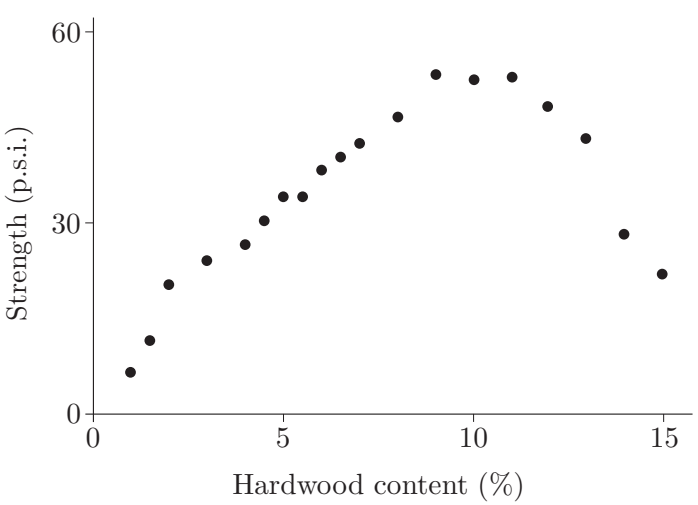


Figure 7 Tensile strength of kraft paper against hardwood content

(a) Which of the two variables, hardwood content and tensile strength, is the response variable and which is the explanatory variable?

(b) What can you say from the scatterplot about the nature of the relationship between the variables?

Notice that in all these examples and activities, the explanatory variable has been plotted along the x-axis in the scatterplot, and the response variable along the y-axis. This is standard practice.

1.2 The general regression model

In regression, it is customary to regard the explanatory variable as non-random and the response variable as a random variable. That is, the values of the explanatory variable are considered 'exact' and hence all the scatter observed in a scatterplot is ascribed to variability in the response. This set-up is directly applicable to the sort of *designed experiments* in which the experimenter is able to choose specific values for the explanatory variable and is interested in the values of the response variable which result. A particular example of this is the duckweed experiment of Example 5; there, the experimenter decided to count the numbers of duckweed fronds (the response variable) at a selected number of values of the explanatory variable, namely after one week, after two weeks, and after each week up to eight weeks. In other regression situations, the values of the explanatory variable might have arisen via some chance mechanism. However, for modelling purposes, interest remains centred on how values of the response variable arise, given those values of the explanatory variables (and not on how the explanatory variables themselves are distributed).

Since the explanatory variable is regarded as non-random, it is always denoted by a lower-case letter, usually x. The response variable is denoted by an upper-case letter, usually Y, to indicate that it is a random variable, whenever it is appropriate to do so. So the points in a general sample of size n are then denoted (x_1, Y_1), (x_2, Y_2), $\ldots$, (x_n, Y_n). That said, the *observed* values of such a sample are usually denoted using lower-case y_is: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n).

In Subsection 1.1, you saw examples where the relationship between two variables appears to be linear and other examples where the relationship might be better modelled by a curve. In each case, there was some scatter about the line or curve – a little in some cases, but a lot in others. The general regression model is made up of two parts:

- (the 'systematic' or deterministic part) a function $h(x)$ that defines the line or curve about which the points in a scatterplot are scattered; $h(x)$ is called the **regression function**
- (the random part) a term which models the scatter, that is, the variation in the response variable about the regression function. This term is itself a random variable, W say. An important property of W is that $E(W) = 0$, that is, that the random part of the model, W, has zero mean.

The general regression model is defined formally as follows.

The general regression model

If the response variable is denoted by Y and the explanatory variable by x, then the **general regression model** for the collection of points (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) can be written

$$Y_i = h(x_i) + W_i, \quad i = 1, 2, \ldots, n.$$

Here h represents some function and the W_is are independent random variables with zero mean.

The function h may be linear, but it can also represent a curve – perhaps polynomial or logarithmic or exponential or trigonometric.

Note that $h(x_i)$ is not random (since x_i is not random) but is an additive constant, so the assumptions for the W_is are equivalent to the response variables Y_i being independent with mean $h(x_i)$. To see the latter, note that

$$\begin{aligned} E(Y_i) &= E\{h(x_i) + W_i\} = E\{h(x_i)\} + E(W_i) = h(x_i) + E(W_i) \\ &= h(x_i) + 0 = h(x_i). \end{aligned}$$

A schematic example of the general regression model is given in Figure 8. There, $h(x) = x^2$ is the regression function which represents the main trend in the model. For each of a number of values of x, the distribution of $Y = h(x) + W = x^2 + W$ is shown; in particular, notice how the distribution is 'centred on' the value $h(x) = x^2$.

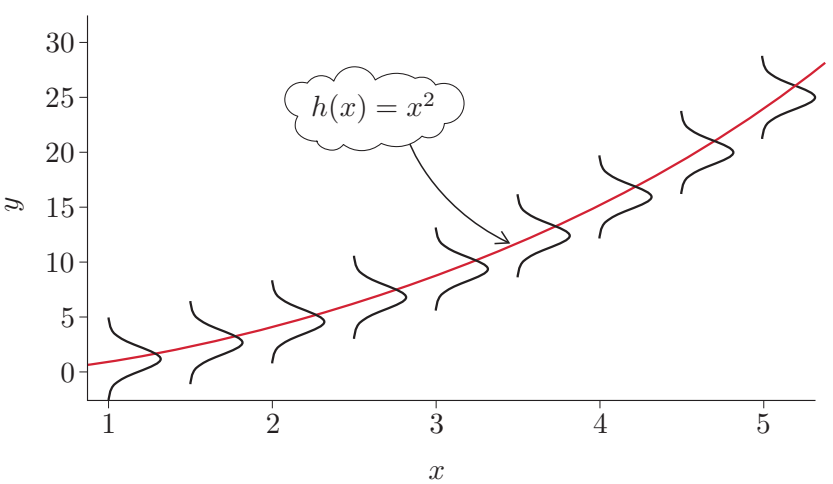


Figure 8 The regression function $h(x) = x^2$ and the distribution of $Y = x^2 + W$ at $x = 1, 1.5, 2, 2.5, \ldots, 5$

Example 6 *A model for Forbes's data*

For Forbes's data in Table 1, the response variable Y is the boiling temperature of water and the explanatory variable x is the atmospheric pressure. From Figure 4, you can see that there appears to be a straight-line relationship between boiling temperature and atmospheric pressure. So a suitable model might be

$$Y_i = \alpha + \beta x_i + W_i.$$

Here α and β are the intercept and slope, respectively, of the straight line relating the boiling temperature to the atmospheric pressure. The random terms W_i account for the scatter around the straight line.

Activity 3 *A model for the heights of Boston schoolboys*

In the Introduction, you saw that the heights of nineteenth-century Boston schoolboys of different ages seemed to be adequately modelled by normal distributions with means linearly related to age and with roughly equal variances. What might be the form of an appropriate regression model for these data? Can you say anything more about the distribution of the random terms in this model?

A little caution is needed here. Sometimes a list of data pairs may appear to suggest a linear relationship between the variables, but when further measurements are taken outside the range investigated, it becomes clear that a more complex model is required. We have already alluded to this in the case of height measurements in the Introduction. There (and in Activity 3 above) a linear relationship was suggested for the mean height of boys between the ages of 6 and 10 years; however, it was noted that such an increasing linear trend would not provide a suitable model for males of older ages. The case of atmospheric pressure and altitude provides another example of this. The scatterplots in Figure 9 show atmospheric pressure (as a percentage of pressure at sea level) plotted against altitude (in metres, at various points on the Earth's surface).

Figure 9 is taken from The Open University (1992) MS284 *An Introduction to Calculus*, Unit 7, *Numbers from Nature*, Milton Keynes, The Open University.

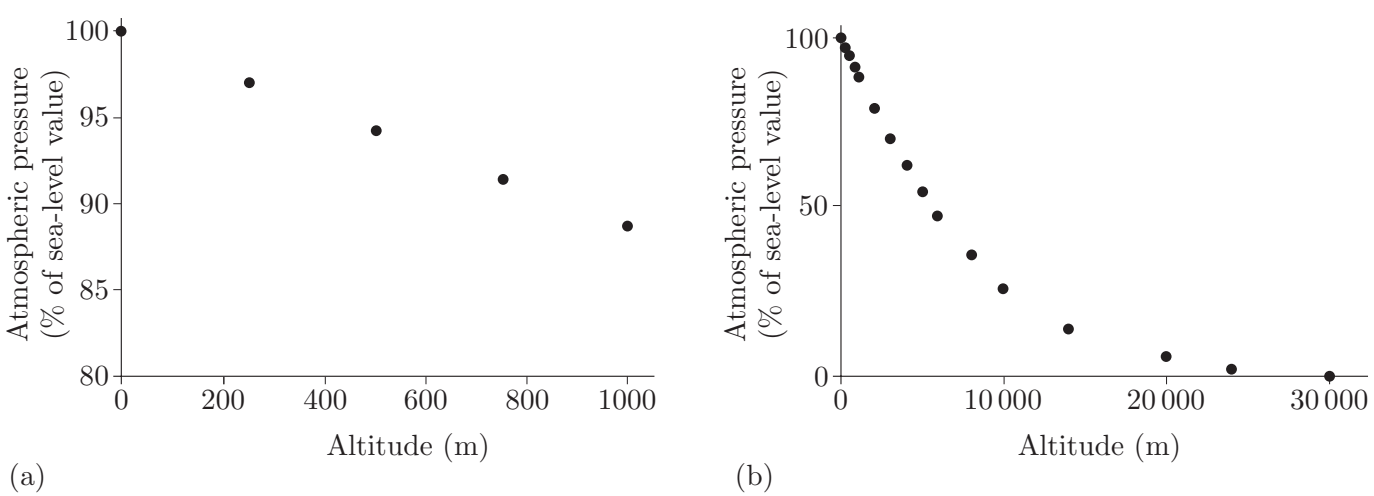


Figure 9 Pressure at different altitudes: (a) up to 1000 metres; (b) up to 30 000 metres

You can see from both panels of Figure 9 that pressure decreases with increasing altitude. Over the range of altitudes considered in Figure 9(a), which was from sea level up to 1000 metres, the relationship appears to be linear. If, however, you were to explore what happens when further

measurements are taken outside this range, you would find that the relationship is no longer linear. This is clear from the scatterplot in Figure 9(b), which shows measured values of atmospheric pressure at altitudes up to 30 000 metres. For this wider range, a more sophisticated mathematical model than a simple straight-line regression model is needed to describe the relationship between the variables.

Example 7 *A model for the duckweed data*

It is clear from Figure 6 that there is a relationship between duckweed growth and passing time, but this relationship is not linear. A possible regression model might be a formula expressing exponential growth, say,

$$Y_i = 20e^{\lambda x_i} + W_i$$

for some parameter value λ. The regression function $h(x) = 20e^{\lambda x}$ is shown for the case $\lambda = 0.5$ in Figure 10. (The value 20 occurs because there were 20 fronds at time 0.) The random term W_i accounts for the scatter. Notice that the regression model of exponential growth cannot persist for all values of the explanatory variable time, else we would now all be covered in duckweed!

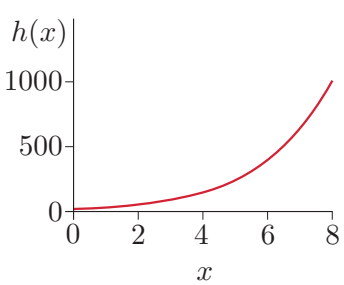


Figure 10 The function $h(x) = 20e^{0.5x}$

1.3 The linear regression model

The most important case of the general regression model is the **linear regression model**. A linear regression model is a regression model where the relationship between Y and x is linear.

The importance of the linear regression model is that not only is it very common (you have already met it in Example 6 and Activity 3), but it is also, as you will see in Section 4, relatively simple to use for statistical inference, such as testing hypotheses and obtaining confidence intervals. In addition, as you will see in Unit 12, particular apparently non-linear regression models can be reduced to linear regression models. A formal definition of the linear regression model is given in the box below.

The linear regression model

If Y is the response variable and x is the explanatory variable, then the **linear regression model** for the collection of points (x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n) can be written

$$Y_i = \alpha + \beta x_i + W_i, \quad i = 1, 2, \ldots, n.$$

The parameters α and β are the intercept and the slope, respectively, of the straight line relating Y to x. The terms W_i are independent random variables with zero mean and constant variance.

In other texts, the random terms W_i are often called random 'errors'. We also use the phrase 'random terms' solely to refer to the W_is in this model, not the Y_is (though they are random, too).

The linear regression model is the special case of the general regression model with the regression function taken to be of the linear form $h(x) = \alpha + \beta x$, together with one additional assumption that it is quite standard to include in the basic linear regression model: the variance of the random term is a constant, $V(W_i) = \sigma^2$ say, for all $i = 1, 2, \ldots, n$.

Activity 4 *The mean and variance of Y_i*

If $E(W_i) = 0$ and $V(W_i) = \sigma^2$, what are $E(Y_i)$ and $V(Y_i)$?

In addition to the results of Activity 4, the response variables Y_i are independent (because the W_is are).

A schematic example of the linear regression model is given in Figure 11 (overleaf). There, $h(x) = 6x - 5$ (the case $\alpha = -5$, $\beta = 6$) is the regression function which represents the main, linear trend in the model. As in Figure 8, the distribution of $Y = h(x) + W$ is also shown for each of a number of values of x.

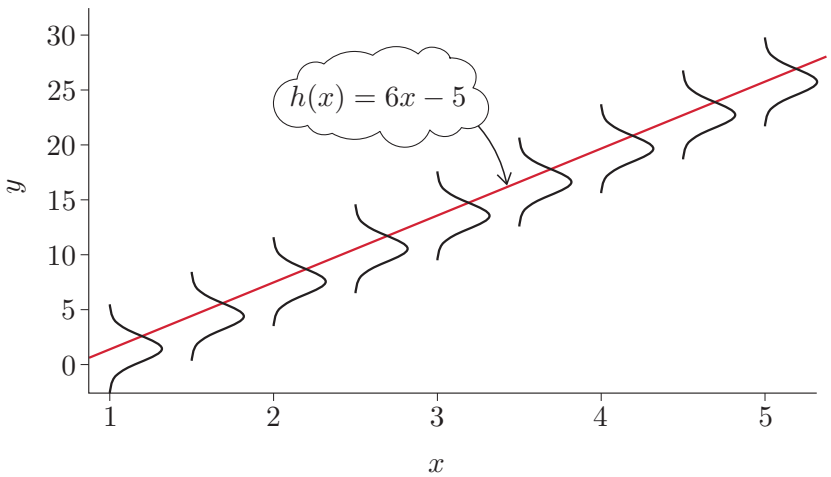


Figure 11 The regression function $h(x) = 6x - 5$ and the distribution of $Y = 6x - 5 + W$ at $x = 1, 1.5, 2, 2.5, \ldots, 5$

The line $y = \alpha + \beta x$ is called the **regression line**. As already mentioned in Example 6 and Activity 3, the parameters α and β can be interpreted as the intercept and slope of the regression line. This interpretation is the same as the usual mathematical one for a straight line. In case you need a reminder:

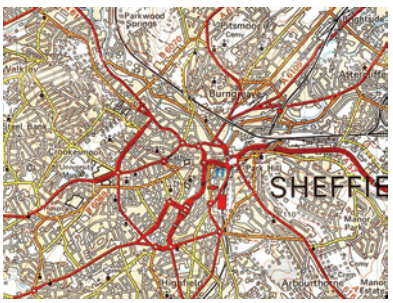
- the intercept α is the value taken by the line when $x = 0$
- the slope β gives how much y changes for every unit change in x. It is also the derivative of the line (for all x). If $\beta > 0$, the line is increasing; if $\beta = 0$, the line is the constant α; and if $\beta < 0$, the line is decreasing.

In the definition of the linear regression model, no assumption has been made about the entire distribution of the random variables W_i or Y_i, just assumptions about their means and variances. However, in order to make inferences, such as testing hypotheses, producing confidence intervals, and so on, it is necessary to assume some distribution for the W_is (or equivalently, for the Y_is). Later in the unit, when inference for linear regression models is discussed, normality of the W_is will be assumed. (You will also learn how to check whether this assumption is reasonable.) If you study statistics further, you will learn about regression models where other distributions are assumed for the W_is or Y_is.

In Example 6 and Activity 3, you saw situations where linear regression models might be useful to describe the relationships between the variables, while a non-linear regression model might be more appropriate for the data in Example 7. Example 8 illustrates a special case of linear regression: the straight line relating Y to x is constrained to go through the origin, that is, the point $x = 0$, $y = 0$.

Example 8 *Distance by road*

Road maps can sometimes be deceptive in the impression they give of distances between two locations. The data in Table 5 are the map distance (that is, the straight-line distance) and the distance by road (both in miles) between twenty different pairs of locations in and around Sheffield. The data raise the following questions. What is the relationship between the two variables? How well can the road distance be predicted from the map distance?



It is clear from the table that the road distance exceeds the map distance in every case. This is hardly surprising: roads tend to have bends, adding to the distance between two points. A scatterplot of road distance against map distance is given in Figure 12.

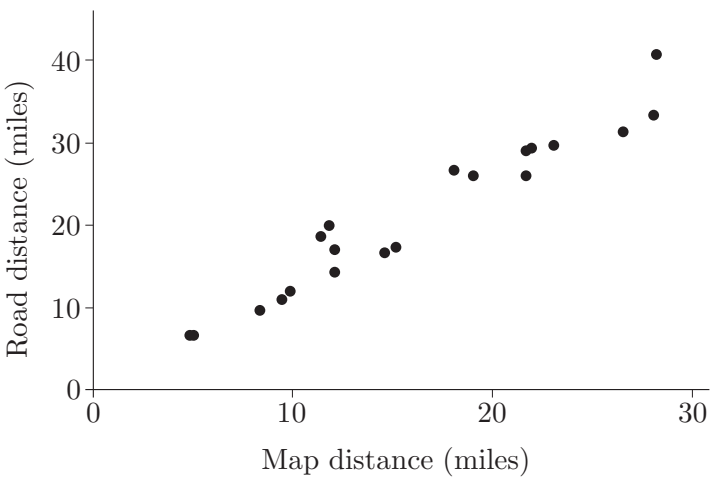


Figure 12 Road distance against map distance between pairs of locations in and around Sheffield

The plot suggests a roughly linear relationship between the two measures. However, the appropriate model here is a little different from that considered in the previous examples. If the map distance between two points is zero (if the two points are the same), then the road distance will also be zero. Therefore the line fitted to the data should go through the origin. That is, the model relating Y (road distance) to x (map distance) should have zero intercept and, since a straight line appears to continue to be a good model all the way to the origin, take the form

$$Y_i = \gamma x_i + W_i.$$

In this model, the parameter γ represents the factor by which a map distance needs to be multiplied to give an estimate of the road distance. The random term W_i again accounts for the scatter identified in the data. Assuming constant variance of the W_is, a linear regression model may be used for the relationship between the variables.

Table 5 Distances in and around Sheffield

Road distance (miles)	Map distance (miles)
10.7	9.5
11.7	9.8
6.5	5.0
25.6	19.0
29.4	23.0
16.3	14.6
17.2	15.2
9.5	8.3
18.4	11.4
28.8	21.6
19.7	11.8
31.2	26.5
16.6	12.1
6.5	4.8
29.0	22.0
25.7	21.7
40.5	28.2
26.5	18.0
14.2	12.1
33.1	28.0

(Source: Gilchrist, W. (1984) *Statistical Modelling*, Chichester, John Wiley and Sons, p. 5)

Notice that the letter γ was used for the slope parameter in the model constrained to go through the origin. It is useful to distinguish in this way between the slopes of the two straight-line models $Y_i = \alpha + \beta x_i + W_i$ and $Y_i = \gamma x_i + W_i$. The constrained model goes by the natural name of **regression through the origin**.

This section concludes with a few points worth noting about linear regression models. First, it is important to realise that it is not necessary to formulate any reason why the relationship between the response variable and the explanatory variable is linear. It is sufficient to argue on the basis of the scatterplot that the relationship *appears* to be linear. Remember also that linearity has been assumed only *within the range* of the data (or just outside the range in Example 8); as mentioned before, you should be cautious about extrapolating outside the range of the data, that is, about assuming that the linearity continues outside the range of the observed data. Finally, you should be aware that statisticians often fit a straight line to data even when there are reasons to believe that something more elaborate is really appropriate. (If you know about Taylor series expansions, you might know that some very complicated curves can be approximated over limited domains by straight lines.)

2 Fitting a linear regression model

In Section 1, you saw several examples of scatterplots where it looked as though a straight-line model would fit the scattered data points (x_i, Y_i) moderately well (in some cases, very well). A practical problem now arises: which straight line fits the data best? In this section, you will see how a technique called the method of least squares can be used to fit the 'best' straight line to the data. The fitted line is called the least squares line.

The method of least squares is discussed in Subsection 2.1. This subsection includes some work using your computer. The special case of a linear regression model where the line is constrained to go through the origin is considered in Subsection 2.2; how to obtain the least squares line for this simple model is described in some detail. In Subsection 2.3, the formula for the least squares line for an 'unconstrained' linear regression model is given without proof.

Fitting a straight line to data by least squares is a method for estimating the parameters α and β of that line. As you will see, the method is quite simple, general and 'natural'. However, you know from Unit 7, in particular, that maximum likelihood is often used to estimate parameters of models from data. So why not use that? Well, in Subsection 2.4, it is shown that if the random terms W_i are from normal distributions, then the slope and the intercept of the least squares line are, in fact, the same as those of the line obtained using the method of maximum likelihood.

2.1 The method of least squares

We begin by looking at a small, illustrative dataset on the way cholesterol level changes with age.

Example 9 *Cholesterol and age*

The data given in Table 6 are the plasma levels (in mg/ml) of total cholesterol in 11 patients aged over 40 who were admitted to a clinic with hyperlipoproteinaemia, a disorder characterised by high levels of lipoproteins in the blood.

Table 6 Cholesterol levels (in mg/ml) and ages (in years)

Age	43	46	48	49	50	52	52	57	57	58	63
Cholesterol	3.8	3.5	4.2	4.0	3.3	4.0	4.3	4.5	4.1	3.9	4.6

(Source: data extracted from a dataset in Krzanowski, W.J. (1998) *An Introduction to Statistical Modelling*, London, Arnold, Chapter 3)

The scatterplot of the data in Figure 13 suggests a roughly linear upward trend but, of course, there is some scatter about the trend due to random variation.

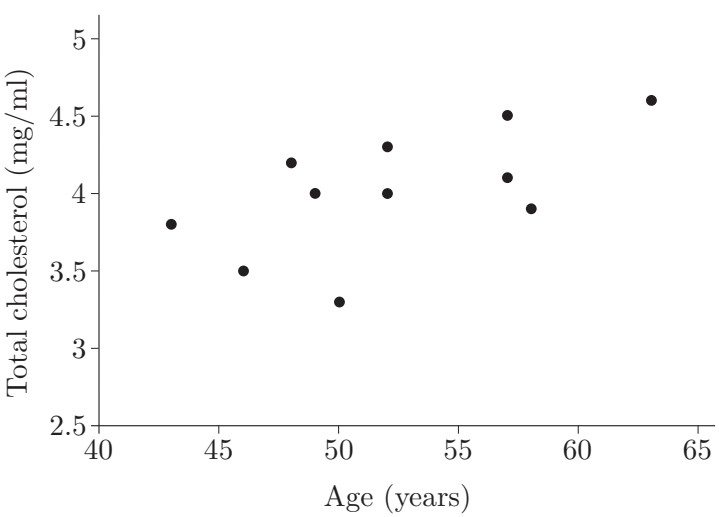
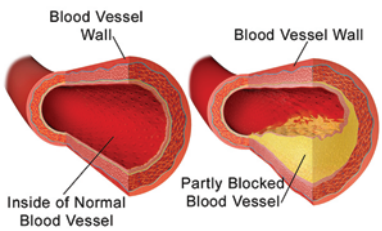


Figure 13 Total cholesterol against age

It seems that a straight-line model of the form

$$Y_i = \alpha + \beta x_i + W_i$$

might describe the data moderately well. Here x_i denotes age (in years), Y_i denotes total cholesterol level (in mg/ml) and W_i is a random term accounting for the scatter. How do we determine the equation of the line which is 'better' than any other line?



Normal and Partly Blocked Blood Vessel

So-called bad cholesterol contributes to plaque, which narrows arteries and can lead to heart disease

This is sometimes more grandiosely called the 'principle of least squares'.

The traditional criterion underlying the estimation of the line that best fits data is the minimisation of a sum of squares of quantities called *residuals*; the resulting method is called the *method of least squares*.

In general, if a line of the form $y = \alpha + \beta x$ is to be fitted to data points (x_i, y_i), then the **residual** w_i for the point (x_i, y_i) is the difference between the observed value y_i and the value of $\alpha + \beta x_i$:

$$w_i = y_i - (\alpha + \beta x_i).$$

The residuals are illustrated for the cholesterol data and one particular choice of line in Figure 14. Notice that the size of each residual is equal to the length of the dashed line joining the data point to the fitted line *vertically* (rather than horizontally or at an angle of 90° to the fitted line).

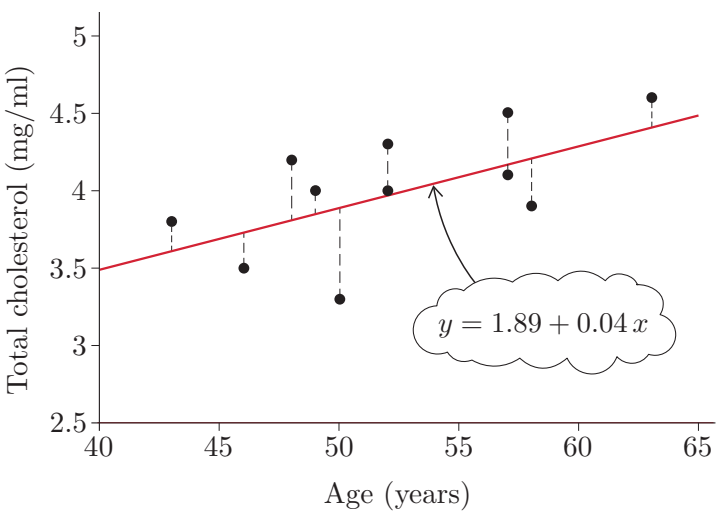


Figure 14 The residuals $w_i = y_i - (\alpha + \beta x_i)$ for one choice of α and β

If the line fits the data well, then the residuals will be small (in absolute value); if not, then at least some of the residuals will be large (in absolute value). When using least squares to choose a best-fitting line, the sum of the squares of the residuals is minimised. The reasoning behind using the sum of *squares* of residuals is the same as that behind summing (and then averaging) squared deviations to form the sample variance (as in Unit 1). You can remind yourself of that reasoning in the following activity.

Activity 5 *Method of least what?*

(a) What is the main reason for choosing the parameters of a regression model to minimise the sum of squared residuals rather than the sum of residuals?

(b) Can you suggest another quantity of the form 'sum of function of residuals' which would have a similar effect to using the function 'square'?

The sum of squared residuals is more often called the **residual sum of squares** and is given by

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2. \quad (1)$$

A small sum indicates a good fit of the line to the data, while a large sum indicates a poor fit. Note that there are two unknown quantities in the expression on the right-hand side of Equation (1): the parameters α and β. The residual sum of squares varies for different values of these parameters. We are interested in the values of α and β that minimise the residual sum of squares (that is, the values that minimise the deviations between the data and the fitted model). The minimising values of α and β are called the **least squares estimates** of the parameters of the regression line, and are denoted by $\widehat{\alpha}$ and $\widehat{\beta}$.

The rest of the work in this subsection consists of a chapter in Computer Book C, in which you can explore the ideas behind the method of least squares.

Refer to Chapter 1 of Computer Book C for the rest of the work in this subsection.

2.2 The least squares line through the origin

Before the formulas for the least squares estimates for the linear regression model are given, a slightly simpler model will be looked at more closely. In Example 8, it was suggested that a good model for the data considered there might be a straight line with the constraint that the line passes through the origin. In this subsection, you will see how to derive the least squares line for this constrained model.

In Example 8, actual road distances between locations in and around Sheffield were compared with direct distances taken from a map. It was decided to fit a straight line passing through the origin to the data. The proposed model was

$$Y_i = \gamma x_i + W_i.$$

A line of the form $y = \gamma x$ has been drawn on the scatterplot of the data in Figure 15 (overleaf) for illustrative purposes only: the value of the slope γ that corresponds to the best straight line through the data is not yet known. The observed residuals based on this line, which are also shown in Figure 15, are in this case given by

$$w_i = y_i - \gamma x_i.$$

Road map or satnav?

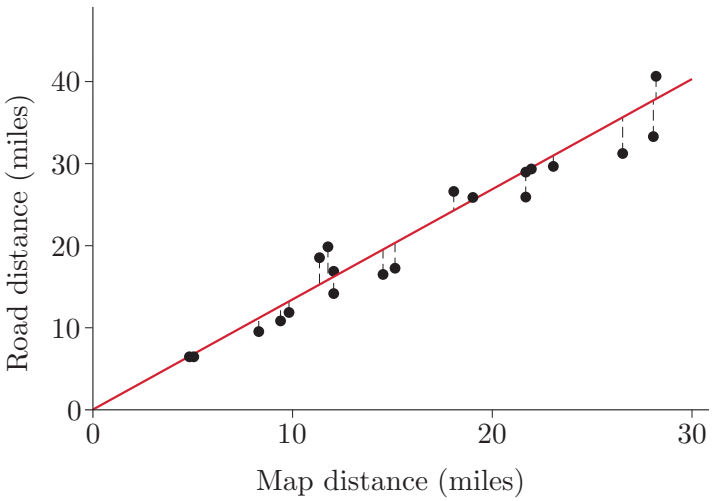


Figure 15 The residuals $w_i = y_i - \gamma x_i$ for one choice of γ

For this model (with the constraint that the straight line goes through the origin), the residual sum of squares is given by

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} (y_i - \gamma x_i)^2 = R(\gamma), \tag{2}$$

say. Here, since $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observed data and therefore are known, there is only one unknown quantity in the residual sum of squares: the slope parameter γ. So the residual sum of squares can be thought of as a function of γ, which we have called $R(\gamma)$. We wish to estimate γ by the value that minimises $R(\gamma)$, that is, that minimises the residual sum of squares. The minimising value of γ is called the *least squares estimate* of the slope of the regression line, and is denoted $\widehat{\gamma}$.

Let us start by taking a look at a graph of $R(\gamma)$. This graph is shown for the Sheffield distance data in Figure 16. ($R(\gamma)$ is plotted only for a limited range of values of γ; for other values of γ, $R(\gamma)$ is even larger and off the scale.) A clear minimum at a value of γ a bit less than 1.3 can be observed.

Now, it turns out that a graph of $R(\gamma)$ always looks very much like the graph of $R(\gamma)$ in Figure 16, whatever the data on which it is based. This is because $R(\gamma)$ is a *quadratic* function of γ, that is, $R(\gamma)$ is of the form $a\gamma^2 + b\gamma + c$ for some coefficients a, b and c.

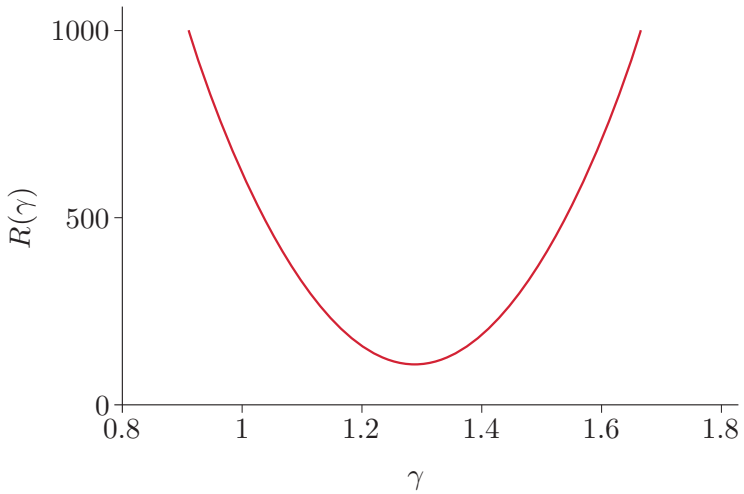


Figure 16 The residual sum of squares

Activity 6 *$R(\gamma)$ as a quadratic function of γ*

By expanding the squared bracket in Equation (2), identify expressions for a, b and c in the representation $R(\gamma) = a\gamma^2 + b\gamma + c$.

In fact, quadratic functions in general look either like that in Figure 16 – 'down-then-up', with a clear minimum – or else like upside-down versions of the function in Figure 16 – 'up-then-down', with a clear maximum. The determinant of shape of a quadratic function is the sign of a. In particular, if $a > 0$, the quadratic function is of the 'down-then-up' variety. (To see this, notice that $a\gamma^2 + b\gamma + c$ necessarily becomes very large as γ becomes very large in absolute value, when $a > 0$.) And it is always the case that $a > 0$ for $R(\gamma)$ because, as you showed in Activity 6, then $a = \sum_{i=1}^{n} x_i^2$ is a sum of squared quantities.

Quadratic curves, or parabolas, abound in the built environment. This one – with $a < 0$! – is the Memorial Cenotaph in Hiroshima Peace Memorial Park, Japan.

Activity 7 *Minimising a quadratic function and hence $R(\gamma)$*

(a) Consider the general quadratic function $ax^2 + bx + c$ with $a > 0$.

(i) Confirm that

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c. \qquad (3)$$

(ii) Hence argue that the minimum of $ax^2 + bx + c$ when $a > 0$ is given by

$$x = -\frac{b}{2a}.$$

(b) By combining the results of Activity 6 and part (a)(ii) above, give an expression for the value of γ that minimises $R(\gamma)$.

For simplicity, the limits $i = 1$ and $i = n$ on the summation symbols have been omitted, and you can do the same from here on.

From the solution to Activity 7(b), the value of γ which minimises the residual sum of squares is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2};$$

and $\widehat{\gamma}$ is the slope of the best straight line through the scattered points that passes through the origin. The equation of the least squares line can be written

$$y = \widehat{\gamma}\, x.$$

These results are summarised in the following box.

The least squares line through the origin

Suppose that it is desired to fit a regression line through the origin and that a scatterplot of data points (x_i, y_i), $i = 1, 2, \ldots, n$, suggests that an appropriate regression model is of the form

$$Y_i = \gamma x_i + W_i,$$

where the W_is are independent with zero mean and constant variance. Then the least squares estimate $\widehat{\gamma}$ of γ is given by

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

The equation of the least squares line through the origin is

$$y = \widehat{\gamma}\, x.$$

Example 10 *Road distances*

In Example 8, Sheffield map and road distances were given. As you have just seen, the least squares estimate of γ in the regression model through the origin depends on two summary statistics, $\sum x_i y_i$ and $\sum x_i^2$. The values of these quantities for the distance data are as follows:

$$\begin{aligned}\sum x_i y_i &= (9.5 \times 10.7) + (9.8 \times 11.7) + \cdots + (28.0 \times 33.1)\\ &= 101.65 + 114.66 + \cdots + 926.80 = 8026.25,\end{aligned}$$

$$\begin{aligned}\sum x_i^2 &= 9.5^2 + 9.8^2 + \cdots + 28.0^2\\ &= 90.25 + 96.04 + \cdots + 784.00 = 6226.38.\end{aligned}$$

So the least squares estimate of the slope γ is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{8026.25}{6226.38} \simeq 1.289.$$

This value corresponds to the exact minimum of $R(\gamma)$ as shown in Figure 16. The equation of the least squares line through the scattered data points is

$$y = 1.289\,x,$$

or, perhaps more intelligibly,

$$\text{road distance} = 1.289 \times \text{map distance}.$$

The least squares line is shown in Figure 17. You can see that the fit is really quite good; the residuals are not large. It seems that the road distance can be predicted quite well from the map distance by multiplying the latter by a factor of 1.289 (that is, by inflating the map distance by a little less than 30%).

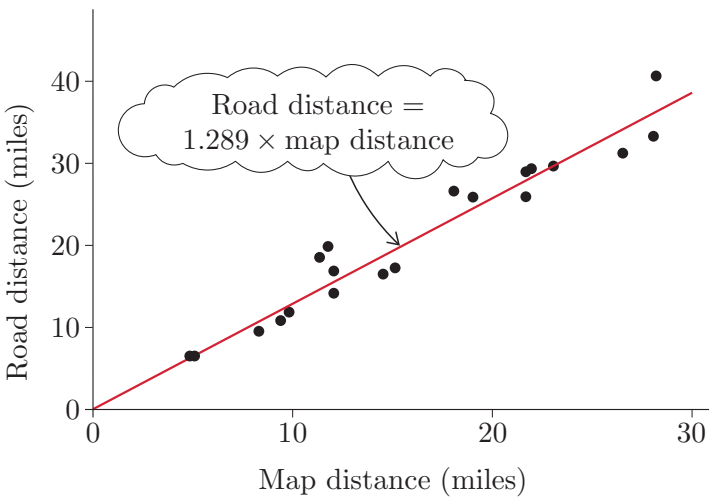


Figure 17 Road distance against map distance, and the least squares line

Activity 8 *Beetles in brackets*

In a botanical experiment, a researcher wanted to estimate the number of a particular species of beetle (*Diaperis maculata*) within fruiting bodies (called brackets) of the birch bracket fungus *Polyporus betulinus*. (This is a shelf fungus that grows on the trunks of dead birch trees.) When the brackets are stored in a laboratory, the beetle larvae within them mature over several weeks. The adults then emerge and can be removed and counted. The bracket weight (in grams) and the number of beetles in each bracket were recorded for a sample of 25 brackets. (Source: Pielou, E.C. (1974) *Population and Community Ecology – Principles and Methods*, New York, Gordon and Breach, pp. 117–21.)

A *Diaperis maculata* fungus beetle

It is suggested that a straight line through the origin might provide an adequate model for the data. The relevant summary statistics for the bracket weight x and the count of beetles y are:

$$\sum x_i^2 = 796\,253, \quad \sum x_i y_i = 219\,817.$$

Calculate the equation of the least squares line through the origin for the data.

2.3 The least squares line

Now consider the 'unconstrained' linear regression model

$$Y_i = \alpha + \beta x_i + W_i.$$

In Subsection 2.1, you saw that the least squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the parameters α and β are the values that minimise the residual sum of squares,

$$\sum (y_i - (\alpha + \beta x_i))^2.$$

There are other routes to the same answer too.

This sum can be minimised using an extension to two parameters of the technique that was used when fitting the least squares line through the origin in Subsection 2.2. However, you will be spared the details. For present purposes it is sufficient simply to write the estimates down. However, before writing them down, it is useful to introduce the following standard shorthand notation.

$$S_{xx} = \sum (x_i - \overline{x})^2 \tag{4}$$

$$S_{yy} = \sum (y_i - \overline{y})^2 \tag{5}$$

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) \tag{6}$$

The expression $(x_i - \overline{x})$ is the deviation of x_i from the mean $\overline{x}$ of the x values, and $(y_i - \overline{y})$ is the deviation of y_i from $\overline{y}$. Thus each term in the sums S_{xx}, S_{yy} and S_{xy} consists of two deviations multiplied together. For this reason S_{xx} and S_{yy} are sometimes called sums of squared deviations, while S_{xy} is a sum of products of deviations. Note that $S_{xx}/(n-1)$ and $S_{yy}/(n-1)$ are the sample variances of the x values and y values, respectively, where n is the sample size.

The easiest way to calculate S_{xx}, S_{yy} and S_{xy} is usually by using the alternative formulas in Equations (7), (8) and (9) below.

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\overline{x}^2 \tag{7}$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = \sum y_i^2 - n\overline{y}^2 \tag{8}$$

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = \sum x_i y_i - n\overline{x}\,\overline{y} \tag{9}$$

That the two versions of each formula within the box immediately above are equal to one another is a simple consequence of recalling that $\overline{x} = \sum x_i/n$ and $\overline{y} = \sum y_i/n$. That Equations (7), (8) and (9) are equivalent to Equations (4), (5) and (6), respectively, takes a bit more algebraic manipulation that you can do for yourself in the next activity.

Activity 9 *Equivalence of formulas*

(a) Check that Equation (7) is equivalent to Equation (4) by manipulating Equation (4).

(b) Why can you now claim that Equation (8) is equivalent to Equation (5) without further mathematical manipulation?

(c) Check that Equation (9) is equivalent to Equation (6) by manipulating Equation (6).

Activity 10 *Calculating S_{xx}, S_{yy} and S_{xy}*

The summary statistics for the cholesterol data from Example 9 are given by

$$n = 11, \quad \sum x_i = 575, \quad \sum y_i = 44.2,$$

$$\sum x_i^2 = 30\,409, \quad \sum y_i^2 = 179.14, \quad \sum x_i y_i = 2324.8.$$

Use Equations (7), (8) and (9) to calculate S_{xx}, S_{yy} and S_{xy}.

We are now ready to write down the formulas for the least squares estimates of the parameters of the linear regression model. First, the least squares estimate $\widehat{\beta}$ of β is given by

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

A similar expression can be written down for $\widehat{\alpha}$, the least squares estimate of α, but it is easier to use the value of $\widehat{\beta}$ to calculate $\widehat{\alpha}$ as follows:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

Then the equation of the least squares line can be written as

$$y = \widehat{\alpha} + \widehat{\beta}\,x.$$

This can be rewritten in various equivalent ways, a popular one being in terms of $\overline{x}$, $\overline{y}$ and $\widehat{\beta}$:

$$y = (\overline{y} - \widehat{\beta}\,\overline{x}) + \widehat{\beta}\,x = \overline{y} + \widehat{\beta}\,(x - \overline{x}).$$

These results may be summarised as follows.

The least squares line

Suppose that the scatterplot of the data points (x_i, y_i), $i = 1, 2, \ldots, n$, suggests that an appropriate regression model might be of the form

$$Y_i = \alpha + \beta x_i + W_i,$$

where the random terms W_i are independent with zero mean and constant variance. Then the least squares estimate $\widehat{\beta}$ of the slope of the regression line is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}},$$

where $S_{xx} = \sum(x_i - \overline{x})^2$ and $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$. The least squares estimate $\widehat{\alpha}$ of the intercept α is given by

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

The equation of the least squares line is

$$y = \widehat{\alpha} + \widehat{\beta}\,x = \overline{y} + \widehat{\beta}(x - \overline{x}).$$

An interesting property of the least squares regression line is given in the next activity.

Activity 11 *Passing through the centroid*

The point on the scatterplot $(\overline{x}, \overline{y})$ is known as the *centroid* of the data. Show that the least squares line passes through the centroid.

Example 11 *Fitting a line to the cholesterol data*

We are now in a position to use least squares to produce a best-fitting line to the cholesterol data discussed in Example 9 and Activity 10.

The summary statistics for the cholesterol data were given in Activity 10. In that activity, you found that $S_{xx} \simeq 352.182$ and $S_{xy} \simeq 14.345$. Hence the least squares estimate of the slope β is

$$\widehat{\beta} \simeq \frac{14.345}{352.182} \simeq 0.04.$$

Using $\widehat{\beta}$ and the summary statistics $n = 11$, $\sum x_i = 575$ and $\sum y_i = 44.2$, the least squares estimate of the intercept α is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{44.2}{11} - \widehat{\beta} \times \frac{575}{11} \simeq 1.89.$$

The value of $\widehat{\beta}$ prior to final rounding has been used in this calculation.

So the equation of the fitted least squares line is

$$y = 1.89 + 0.04\,x.$$

Alternatively, the model can be written as

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age},$$

where total cholesterol is measured in mg/ml and age is given in years. The least squares line is shown in Figure 18. (It is the same line as was shown in Figure 14.) The line appears to fit the data reasonably well.

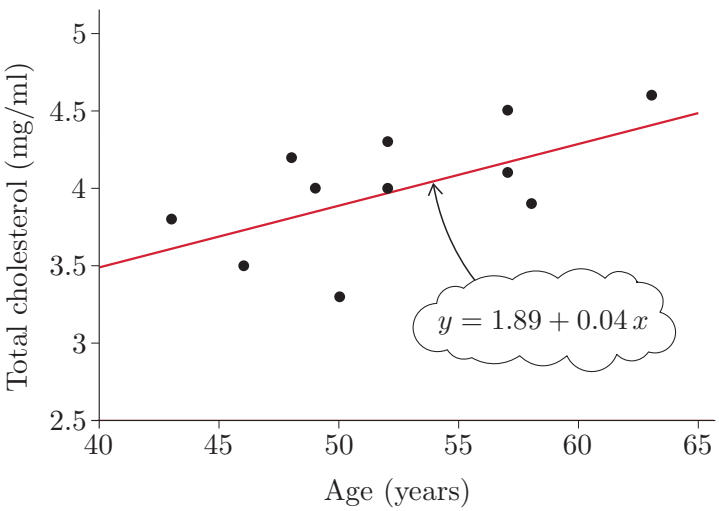


Figure 18 Total cholesterol against age, and the least squares line

In terms of interpretation, the estimated value of the intercept, $\widehat{\alpha}$, is of little interest in this particular context because it refers to a person of age 0 years, whereas the linear model is fitted – and assumed appropriate – to data on people over the age of 40 years. The estimated value of the slope, $\widehat{\beta} \simeq 0.04$, is of interest, however. It tells us that, for patients with hyperlipoproteinaemia aged over 40 years, an increase in age of one year is expected to lead, on average, to an increase in total cholesterol of about 0.04 mg/ml.

Another use of the least squares fitted regression line is for *prediction*. Suppose that another individual of the same type as those to whom the line was fitted has a value x_0 say, of the explanatory variable; however, we do not yet know the value of the response variable, y_0 say, for this individual. The least squares line allows us to predict what we think that value might be, by setting $x = x_0$ in the equation of the least squares line:

$$y_0 = \widehat{\alpha} + \widehat{\beta}\, x_0.$$

Prediction via statistics or crystal ball?

Example 12 *Predicting total cholesterol*

As an example, the least squares line obtained in Example 11, $y = 1.89 + 0.04\,x$, could be used to predict the total cholesterol level of a person with hyperlipoproteinaemia aged over 40. For example, for someone aged 45, the fitted line predicts a total cholesterol level of

$$1.89 + 0.04 \times 45 = 3.69 \text{ mg/ml},$$

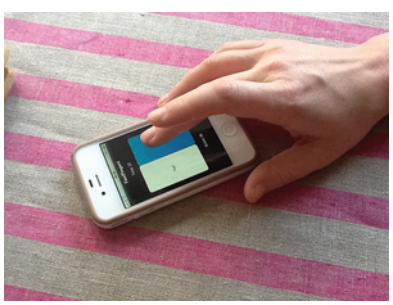
while for someone aged 60, the fitted line predicts a total cholesterol level of

$$1.89 + 0.04 \times 60 = 4.29 \text{ mg/ml}.$$

Notice, however, that these are single-value or 'point' predictions, without any indication of uncertainty concerning that prediction. We are not claiming that, for example, in Example 12, everyone with hyperlipoproteinaemia aged 60 should have a cholesterol value of exactly 4.29 mg/ml, just that 4.29 mg/ml seems to be a reasonable prediction of the average value of cholesterol for people with this condition aged 60. Indeed, any prediction of the form $y_0 = \widehat{\alpha} + \widehat{\beta}\, x_0$ is actually an estimate of the average value, $\alpha + \beta\, x_0$, of the response for an individual with $x = x_0$. As point estimates have corresponding interval estimates (Unit 8), so point predictions have corresponding interval predictions; these will be considered briefly in Subsection 4.3 to follow.

Activity 12 *Finger-tapping*

Finger-tapping is a fairly standard psychological task performed by subjects to assess alertness through manual dexterity.



Modern finger-tap testing

An experiment was carried out to investigate the effect of the stimulant caffeine on performance on a simple physical task. Thirty male college students were trained in finger-tapping; it is the speed of finger-tapping that is of interest. They were then randomly divided into three groups of ten, and the students in each group received different doses of caffeine (0 mg, 100 mg and 200 mg). Two hours after treatment, each student was required to do finger-tapping, and the number of taps achieved per minute was recorded. The recorded figures for each of the 30 students are given in Table 7.

Table 7 Finger-tapping

Caffeine dose (mg)	Taps per minute									
0	242	245	244	248	247	248	242	244	246	242
100	248	246	245	247	248	250	247	246	243	244
200	246	248	250	252	248	250	246	248	245	250

(Source: Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425)

It is not possible to deduce very much about the shape of the variation in tapping performances at each dose level from the scatterplot shown in Figure 19. However, there is some evidence of a linear upward trend.

For clarity, coincident points are shown slightly displaced, or 'jittered', vertically in Figure 19.

Suppose that we wish to model the relationship between tapping performance Y and caffeine dose x by a linear regression model. The summary statistics for the data in Table 7 are given by

$$n = 30, \quad \sum x_i = 3000, \quad \sum y_i = 7395,$$

$$\sum x_i^2 = 500\,000, \quad \sum x_i y_i = 743\,000.$$

(a) Use the summary statistics to calculate S_{xx} and S_{xy}.

(b) Calculate the equation of the least squares line for the data.

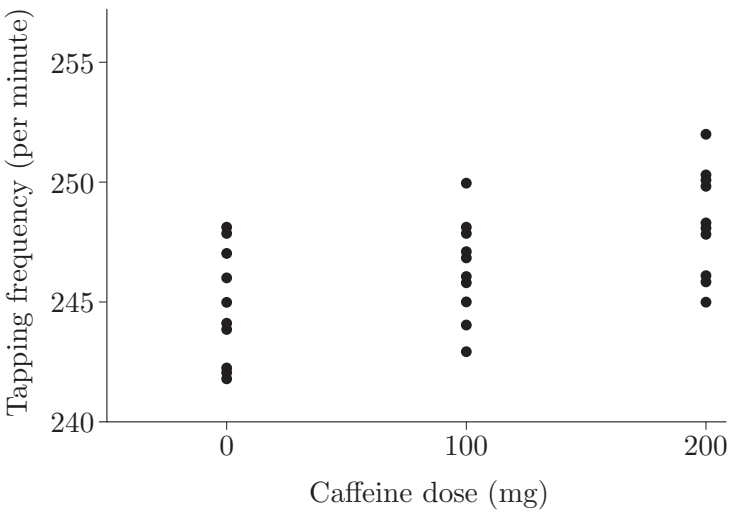


Figure 19 Tapping performance against caffeine dose

(c) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(d) Use the equation of the fitted least squares line to predict the number of taps per minute of a student treated with 50 mg of caffeine.

In practice, a computer is almost always used to fit least squares lines. You will do this using Minitab in Section 3.

2.4 Maximum likelihood estimation in regression

In this subsection, you will see that if normality is assumed for the random terms W_i in a linear regression model, then the least squares estimates of the parameters of the line are also the maximum likelihood estimates of those parameters. This argument further justifies the use of least squares estimation in regression. The argument is given in full for completeness and worked through in Screencast 11.1. If you cannot follow all the details, don't worry: the result is worth knowing but you won't need to be able to reproduce the argument leading to it.

Suppose that the random terms W_i, $i = 1, 2, \ldots, n$, come from independent normal distributions, that is,

$$W_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$

Notice that each of these normal distributions has zero mean and the same variance, σ^2, these being general properties of the random terms in the linear regression model. Equivalently, since $Y_i = \alpha + \beta x_i + W_i$ and the $\alpha + \beta x_i$ terms can be treated as constants,

See Activity 4 in this unit and Distributional Result (3) of Unit 6.

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, 2, \ldots, n,$$

and these normal distributions are independent also.

This is just for simplicity: the least squares estimates of α and β are the maximum likelihood estimates of α and β when σ^2 is not known too.

Suppose for the remainder of this subsection that the value of σ^2 is known. Then, following the discussion of likelihood estimation for continuous distributions in Unit 7 (with two unknown parameters α and β replacing the single unknown parameter θ there), the likelihood in this case is

$$L(\alpha, \beta) = f(y_1; \alpha, \beta) \times f(y_2; \alpha, \beta) \times \cdots \times f(y_n; \alpha, \beta),$$

where $f(y_i; \alpha, \beta)$ is the p.d.f. of Y_i when $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. Using the p.d.f. of the normal distribution from Unit 6, the likelihood is therefore

$$\begin{aligned} L(\alpha, \beta) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_1 - (\alpha + \beta x_1)}{\sigma}\right)^2\right] \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_2 - (\alpha + \beta x_2)}{\sigma}\right)^2\right] \\ &\qquad \vdots \\ &\quad \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_n - (\alpha + \beta x_n)}{\sigma}\right)^2\right] \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2\right]. \end{aligned}$$

Now, the maximum likelihood estimates of α and β (when σ^2 is known) are the values that maximise the likelihood. The first term in the likelihood is a constant (since σ^2 is known) and the second term is of the form

$$\exp\{-kR(\alpha, \beta)\},$$

where $k = 1/(2\sigma^2)$ is a positive constant and

$$R(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2.$$

Since the function e^{-kx}, $k > 0$, is decreasing (see, for example, Figure 2(a) of Unit 5 or Figure 13 of Unit 8), the second term in the likelihood (and hence the whole product) is maximised with respect to α and β when the quantity $R(\alpha, \beta)$ is minimised. However, $R(\alpha, \beta)$ can be recognised as precisely the residual sum of squares, given in Equation (1), that is minimised to find the least squares estimates.

So, under the assumption of normality with known variance, the least squares estimates of α and β are the same as the maximum likelihood estimates of α and β.

The above argument is reviewed in Screencast 11.1.

Screencast 11.1 Maximum likelihood estimation of regression parameters assuming normality is the same as least squares estimation

Exercise on Section 2

Exercise 1 *The least squares line for Forbes's data*

For Forbes's data, which are given in Table 1, the summary statistics are as follows:

$$n = 17, \quad \sum x_i = 426, \quad \sum y_i = 3450.2,$$

$$\sum x_i^2 = 10\,820.9966, \quad \sum x_i y_i = 86\,735.495.$$

(a) Use the summary statistics to calculate the equation of the least squares line for the data.

(b) Interpret what the values of the least squares estimates of the parameters of the regression line tell us.

(c) Use the fitted line to obtain a point prediction of the boiling point of water at an atmospheric pressure of 25 inches Hg.

3 Checking the assumptions

You will shortly be using Minitab to fit linear regression models to some of the datasets described in Sections 1 and 2. Before doing that, however, there is an important question to ask that was neglected in Section 2: how can we check that a fitted model is reasonable? For the linear regression model

$$Y_i = \alpha + \beta x_i + W_i,$$

the basic assumptions are as follows.

1 The random terms W_i are independent.

2 The W_is have zero mean and constant variance σ^2.

It's always worth checking your assumptions about the British weather

Remember that, as you showed in Activity 4, the assumption of zero mean for the W_is is equivalent to the assumption that a line of the form $\alpha + \beta x$ is appropriate for the mean of the Y_is:

$$E(Y_i) = \alpha + \beta x_i.$$

So, together, the two basic assumptions of linear regression are that the W_is are independent random variables with zero mean and variance σ^2. Or, equivalently, the two basic assumptions of linear regression are that the Y_is are independent random variables with mean $\alpha + \beta x_i$ and variance σ^2.

Assumption 1 on independence of the $W_i s$ can be checked, although it will not be done here: independence has to do with the design of the experiment and how the data were collected. It will usually be clear whether or not the independence assumption is justifiable.

Assumption 2, that the W_is have zero mean and constant variance, can be checked using a diagram called a *residual plot*. This is the topic of the next subsection.

3.1 Residual plots

We wish to check the properties of the W_is. Well, the linear regression model can be rearranged as

$$W_i = Y_i - (\alpha + \beta x_i).$$

However, because α and β are unknown, we immediately come up against the problem that the W_is cannot be observed. The W_is can, though, be estimated in the natural way via the estimated values of $\alpha + \beta x_i$. The latter are

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i,$$

Fitted values differ from predicted values by their values of x: fitted values are for observed values $x = x_i$ while predicted values are for other values $x = x_0$.

which we will now refer to by the standard nomenclature of **fitted values**. The required estimates of the W_is are therefore the differences between the observed values, y_i, and the fitted values, $\widehat{y}_i$, namely the quantities

$$w_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i).$$

And these you have seen before: the w_is are, in the terminology of Subsection 2.1, residuals – specifically residuals from the least squares fitted line (but they will be referred to just as residuals from now on). So the modelling assumptions can be checked by looking to see whether the residuals w_i, used as estimates of the random terms W_i, might have come from some distribution with zero mean and variance σ^2 (for some value of σ^2).

Elsewhere, a residual plot is sometimes defined to be a scatterplot of the observed residuals w_i against the values x_i of the explanatory variable. For the linear regression model discussed in this section, the two plots are the same after rescaling.

In this module, a **residual plot** is defined to be a scatterplot of the observed residuals w_i against the fitted values $\widehat{y}_i$. If Assumption 2 is satisfied, then the residuals should be scattered about zero in a random, unpatterned fashion. Note that the residuals are the deviations from the fitted model: a pattern in the residual plot would suggest a dependence between the residuals and the corresponding fitted values, indicating a breach of the assumption that the random terms W_i, which the residuals w_i are estimating, have zero mean and constant variance. The key here is actually that the mean and variance of the random terms should both be *constant*. Patterns in the residuals when plotted against the fitted values suggest that either the mean or the variance or both are not constant. Figure 20 shows four typical shapes of residual plots.

Figure 20(a) is a residual plot with no apparent pattern of any kind in the residuals: this is the type of plot that you might expect to obtain when the assumptions are justified. There is a definite pattern in each of the other panels of Figure 20.

In Figure 20(b), when moving from left to right, from smaller to larger fitted values, the residuals are negative at first, then positive, then negative again. In general, a residual plot displaying a pattern such as this is an indication that the assumption of constant, zero mean may not be

valid – that is, that the relationship between the response and explanatory variables is not linear. (The pattern in the residuals gives an indication of what the regression function should have been – perhaps, in this instance, quadratic rather than linear.)

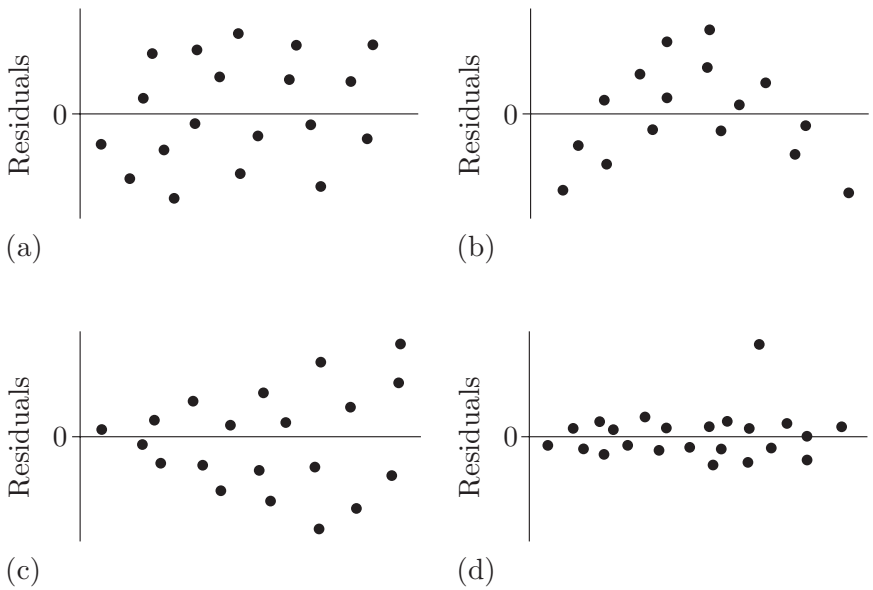


Figure 20 Residual patterns, when plotted against fitted values: (a) unpatterned, (b) a systematic discrepancy, (c) variance not constant, (d) an outlier

In Figure 20(c), the pattern is indicative of the variance of the random terms not being constant: the variability of the residuals (and hence presumably the variance of the random terms) increases as the fitted values increase. (In variations on this theme, the variance of the residuals might decrease as the fitted values increase, or even exhibit some other pattern, such as small variance then larger variance then smaller variance again.)

Finally, the residual plot in Figure 20(d) has a single residual that is considerably larger in magnitude than any of the others. The plotted point may correspond to an outlier.

Let's see what residual plots can tell us in an example and a couple of activities.

Example 13 *Checking residuals for the cholesterol data*

In Example 11, the least squares line was fitted to the cholesterol data. The equation of the line is

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}.$$

This line was shown on a scatterplot of the data in Figure 18. The figure is repeated in Figure 21 (overleaf).

This figure is not a residual plot!

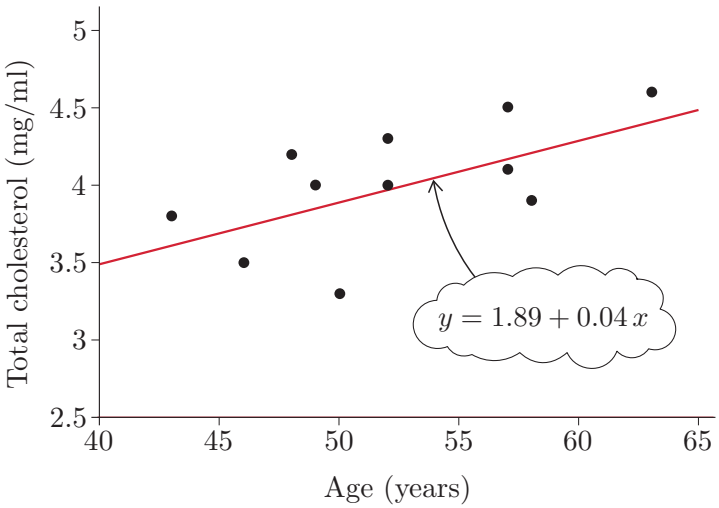


Figure 21 Total cholesterol against age, and the least squares line

In order to check Assumption 2, it is necessary to calculate the fitted values and the residuals. In this case, for each age $x_1, x_2, \ldots, x_n$, the fitted values $\widehat{y}_i$ are given by

$$\widehat{y}_i = 1.89 + 0.04\,x_i$$

and the residuals w_i are then found from

$$w_i = y_i - \widehat{y}_i = y_i - 1.89 - 0.04\,x_i.$$

A residual plot for the cholesterol data, that is, a scatterplot of the residuals w_i against the fitted values $\widehat{y}_i$, is shown in Figure 22.

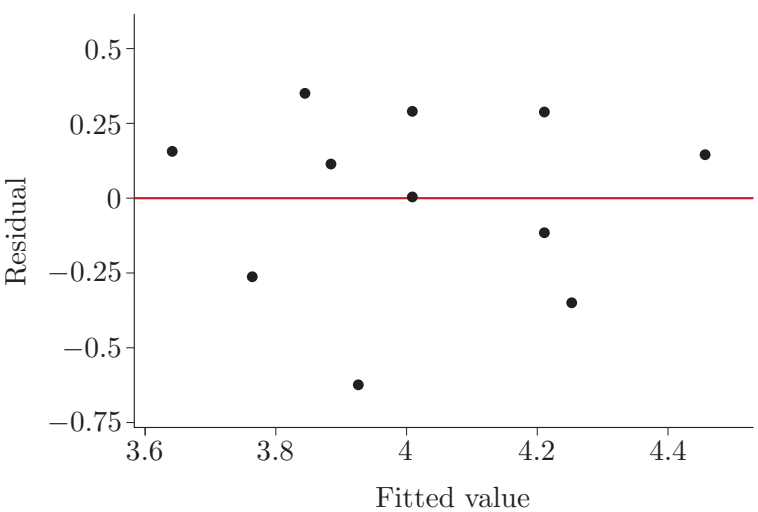


Figure 22 A residual plot for the cholesterol data

This plot shows no particular pattern, and there are approximately the same number of points below the line as above it; the points seem to be randomly scattered around zero. That is, Assumption 2 seems to be satisfied. So a linear regression model might provide an adequate model for these data.

Activity 13 *Checking residuals for Forbes's data*

Forbes's data on the way in which the boiling point of water depends on pressure were introduced in Example 3. The data were shown in Figure 4, and the linear regression model was fitted to the data by least squares in Exercise 1. The equation of the least squares fitted line is

$$y = 155.30 + 1.90\,x,$$

where y is the boiling point of water in °F and x is the pressure in inches of mercury.

If the modelling assumptions are reasonable for these data, then the W_is are observations with a constant, zero mean and an unknown but constant variance. A residual plot is given for this fitted regression line in Figure 23.

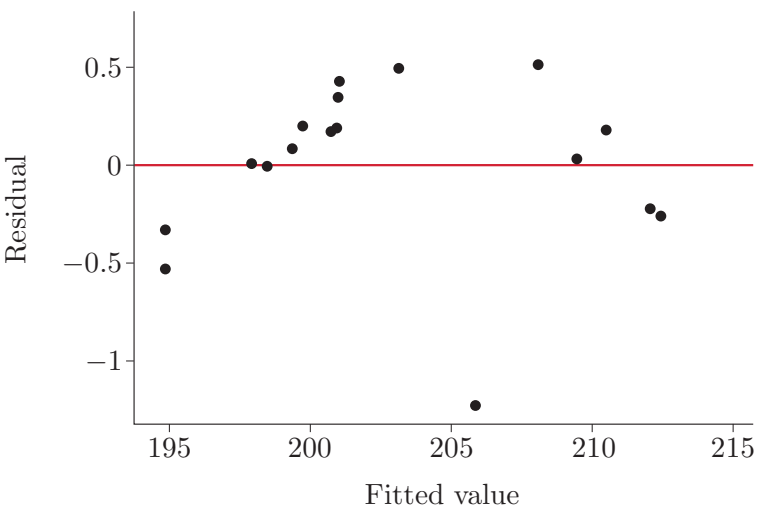


Figure 23 A residual plot for Forbes's data

Comment on what this plot tells you. Is the linear regression model a good one for these data?

Activity 14 *Defects in the Trans-Alaska oil pipeline*

In Subsection 5.4 of Unit 1, we looked at a dataset of size $n = 107$ concerning the measurement of defects in the Trans-Alaska oil pipeline. Depths of defects were measured in the field using ultrasonic measuring equipment and again, potentially more accurately, in the laboratory later. Interest now centres on how well calibrated in-field measurements of pipeline defects were, in the sense of how closely they depend on their corresponding laboratory measurements.

Ultrasonic pipeline inspection

In Example 26 of Unit 1, with the help of a scatterplot interpretation checklist, we decided that the data exhibit a 'moderately strong' positive linear relationship with no obvious outliers. The data therefore seem ripe

for modelling by linear regression. This was done and the fitted least squares line turned out to be

$$\text{field defect depth} = 4.99 + 0.731 \times \text{laboratory defect depth}.$$

The data together with the least squares line are plotted in Figure 24. The line appears to fit the data pretty well.

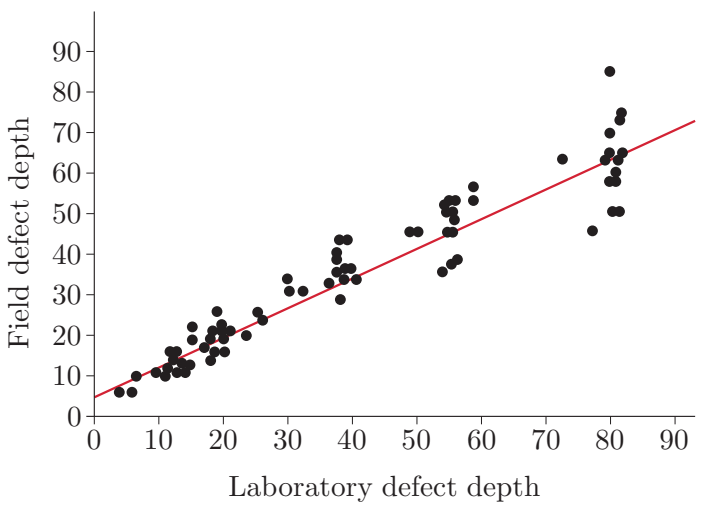


Figure 24 The Trans-Alaska oil pipeline data and least squares line

But what of the wider linear regression model with its assumption of constant, zero mean and constant variance of the random terms W_i? Are these assumptions (collectively Assumption 2) justified by the residual plot provided in Figure 25? If not, why not?

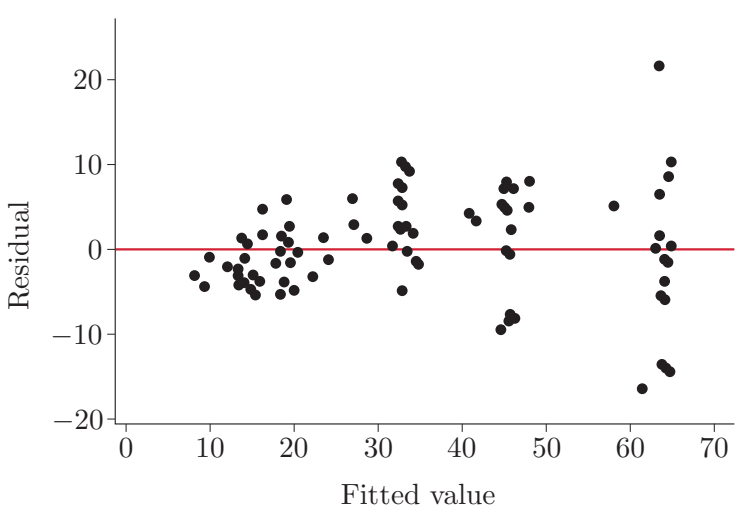


Figure 25 Residual plot for the Trans-Alaska oil pipeline data

A small point to note is that for the linear regression model including both slope and intercept terms (i.e. not regression through the origin), the sum

of the residuals is always zero. If a plot purported to be a residual plot clearly violates this property, then something has gone wrong in producing that residual plot. You can prove this property of residuals for yourself in the next activity.

Activity 15 *Summing the residuals*

The residuals w_i can be written $w_i = y_i - \widehat{\alpha} - \widehat{\beta}\,x_i$ where $\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}$. Use these facts to show that $\sum_{i=1}^{n} w_i = 0$.

3.2 Checking normality of residuals

In order to use the fitted regression model to make inferences, test hypotheses, produce confidence intervals, and so on, it is necessary to assume some distribution for the W_is. The most common assumption to make is the one made in Subsection 2.4: that the random terms are normally distributed. Sometimes other distributions are used – for example, the Poisson distribution or the Bernoulli distribution, where appropriate; you may well come across some of these in further statistical studies. However, for inferential work on the linear regression model in this module, the following assumption will be made.

Using these distributions requires further modifications to the linear regression model, however.

3 The W_is are normally distributed.

If Assumptions 1 to 3 are satisfied, then the W_is are independent normal random variables with zero mean and some variance σ^2. That is, the W_is can be regarded as a random sample from an $N(0, \sigma^2)$ distribution, and the Y_is can be regarded as a random sample from an $N(\alpha + \beta\,x_i, \sigma^2)$ distribution (as in Subsection 2.4).

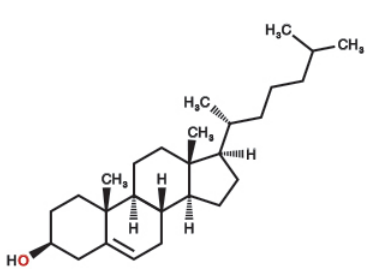
A normal probability plot can be used to check whether it is reasonable to assume that a sample of data comes from a normal distribution (Section 5 of Unit 6). So, if a residual plot has confirmed that a linear regression model is appropriate for the data at hand, in the sense that we can assume that the W_is have zero mean and constant variance, then Assumption 3 can be checked using a normal probability plot of the residuals. This is illustrated in Example 14 and Activity 16 to follow.

Example 14 *Normality of residuals for the cholesterol data*

The cholesterol data introduced in Example 9 were fitted in Example 11 by the least squares regression line

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}.$$

In Example 13, it was concluded, on the basis of the residual plot in Figure 22, that it was reasonable to make Assumption 2 (zero mean, constant variance of W_is).



The structure of a cholesterol molecule

In order to now check Assumption 3 (normality of W_is), a normal probability plot of the residuals w_i is shown in Figure 26.

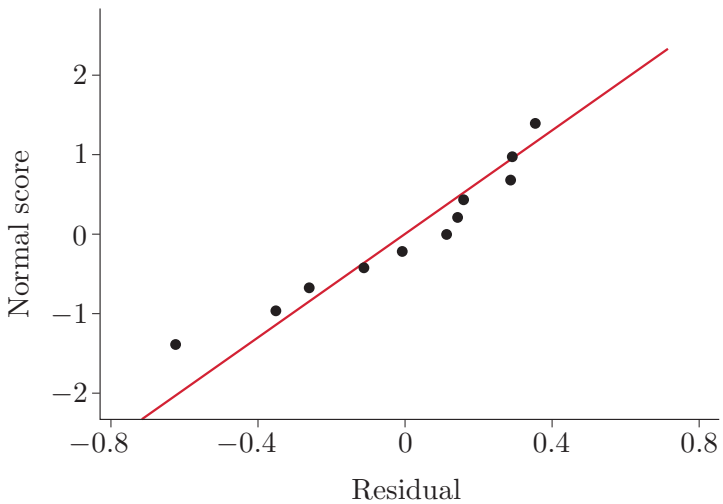


Figure 26 A normal probability plot of the residuals for the cholesterol data

The residuals lie reasonably close to a straight line, so it seems plausible that the random terms come from a normal distribution. That is, it can be argued that Assumption 3 seems to be reasonable for these data. (You might, however, perceive a curve in the probability plot, but with so few data points it seems insufficiently strong to rule out the normality of random terms in the model.)

Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of total cholesterol on age for patients aged over 40 years with hyperlipoproteinaemia.

Activity 16 *Checking the assumptions for the tapping data*

The tapping data introduced in Activity 12 were fitted there by the least squares regression line

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Here, the response variable is the number of taps per minute and the explanatory variable, caffeine dose, is measured in mg. In order to check the assumptions of the linear regression model, a residual plot and a normal probability plot of residuals are given in Figure 27(a) and Figure 27(b), respectively. Similarly to the scatterplot of the original data in Figure 19, coincident points in the residual plot are shown jittered slightly, in the vertical direction.

Comment on what these plots tell you. Is the linear regression model a good one for these data?

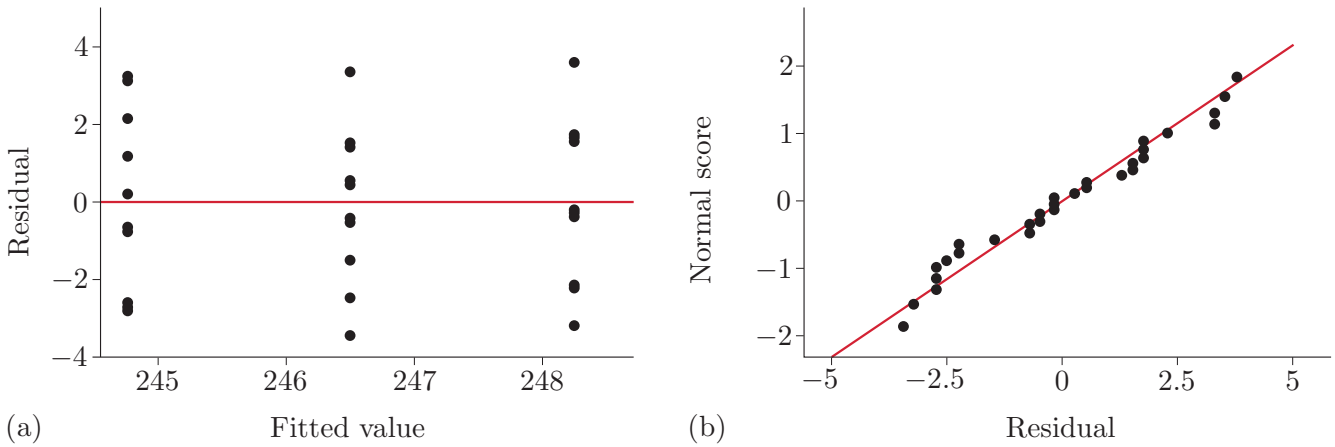


Figure 27 Plotting the residuals for the tapping data: (a) residual plot; (b) normal probability plot

A computer software package is usually used to produce residual plots and normal probability plots of residuals. However, although a computer can do the calculations and draw the plots, it is up to you to interpret the plots and assess whether the assumptions are reasonable! So now go to Computer Book C. There you will use Minitab to fit linear regression models to data and to check the assumptions of the fitted models.

Refer to Chapter 2 of Computer Book C for the rest of the work in this section.

It should be added that all is not lost if the assumptions of the linear regression model are not met. Further regression modelling of the data can still be performed. One way of accounting for failures in the assumptions will be investigated in Unit 12.

Exercise on Section 3

Exercise 2 *Cholesterol and a wider range of ages*

The data given in Table 8 are the plasma levels (in mg/ml) of total cholesterol in 24 adults with hyperlipoproteinaemia. This is the full dataset from which the smaller dataset studied so far in this unit was extracted. The smaller dataset concerned only those individuals aged over 40 years; the full dataset adds 13 other patients who were aged 40 years or under.

Table 8 Cholesterol levels (in mg/ml) and ages (in years)

Age	20	22	22	24	25	28	28	29	30	33	34	36
Cholesterol	1.9	2.1	2.5	2.5	3.0	2.3	2.9	3.3	2.6	3.0	3.2	3.8
Age	40	43	46	48	49	50	52	52	57	57	58	63
Cholesterol	3.2	3.8	3.5	4.2	4.0	3.3	4.0	4.3	4.5	4.1	3.9	4.6

(Source: full dataset from Krzanowski, W.J. (1998) *An Introduction to Statistical Modelling*, London, Arnold, Chapter 3)

For the data concerning the over-40s only, the least squares line

$$\text{total cholesterol} = 1.89 + 0.04 \times \text{age}$$

was fitted in Example 11. In Examples 13 and 14, Assumptions 2 and 3 were checked. In Example 14, it was concluded that: 'Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of total cholesterol on age for patients aged over 40 years with hyperlipoproteinaemia.' This exercise concerns the question of whether or not the linear regression model with normally distributed random terms remains appropriate to explain the dependence of total cholesterol on age for all adult patients with hyperlipoproteinaemia.

(a) A scatterplot of the full dataset is provided in Figure 28 along with the least squares line fitted to these data. On the basis of this plot, does there seem to be a case for use of the linear regression model?

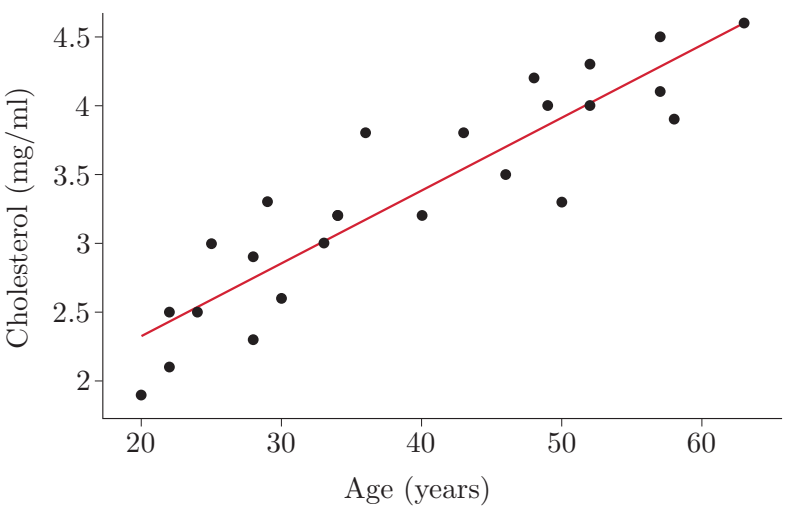


Figure 28 Total cholesterol against age, and the least squares line, for the full dataset

(b) A residual plot for the data in Figure 28 is provided in Figure 29. Does Assumption 2, that the random terms have constant, zero mean and constant variance, seem to be satisfied?

(c) A normal probability plot of the residuals for the data in Figure 28 is provided in Figure 30. Does Assumption 3, that the random terms are normally distributed, seem to be satisfied?

(d) The least squares line in Figure 28 has the formula

$$\text{total cholesterol} = 1.28 + 0.05 \times \text{age}.$$

How, numerically, do the least squares lines fitted to the different versions of the dataset differ? Sketch the lines on a graph as functions of age (in years): for the whole dataset on the range 20 to 63, and for the cut-down dataset on the range 43 to 63. Does the line appear to have changed much since the younger patients' data were included?

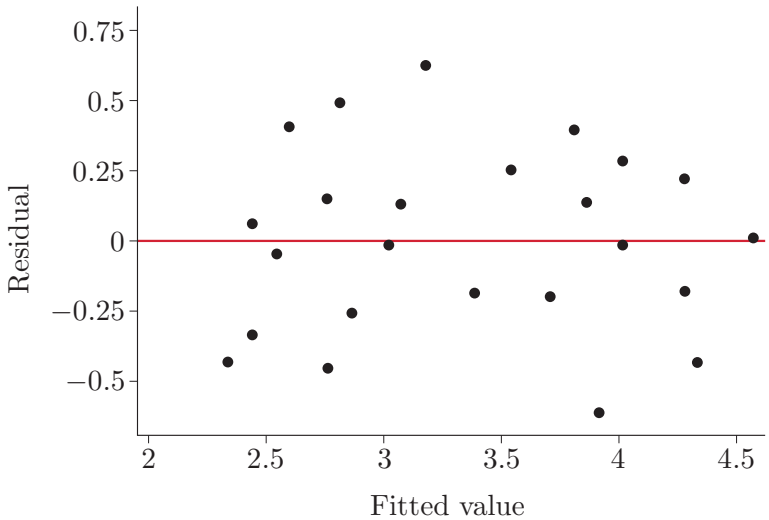


Figure 29 A residual plot for the full cholesterol dataset

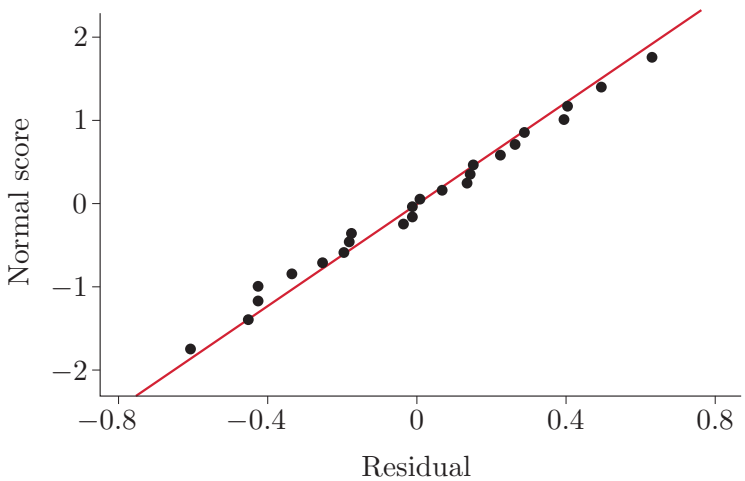


Figure 30 A normal probability plot of the residuals for the full cholesterol dataset

4 Sampling properties and statistical inference

In Section 2, you learned how to calculate the least squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the parameters α and β of the regression line. However, a repeated experiment would almost certainly result in different responses and hence different estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β. The estimates $\widehat{\alpha}$ and $\widehat{\beta}$ vary from one experiment to the next, so they are observations of random

Results similar to the results in this section are available for the estimator $\widehat{\gamma}$ of the constrained model. They will not be given in this module.

variables. It is standard to use the same notation $\widehat{\alpha}$ and $\widehat{\beta}$ for these random variables as for the individual estimates. These random variables are called the **least squares estimators** of α and β. The formula for the least squares estimator of β is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

In this case, remembering that the explanatory variable is regarded as non-random in regression, S_{xy} is the random variable given by

$$S_{xy} = \sum (x_i - \overline{x})(Y_i - \overline{Y}).$$

Here, $\overline{Y}$ denotes the mean of the random variables $Y_1, Y_2, \ldots, Y_n$, that is, $\overline{Y} = \sum Y_i/n$.

Notice that the same notation S_{xy} is used for this sum, which involves random variables, as was used in Subsection 2.3 for the sum of products of deviations $\sum(x_i - \overline{x})(y_i - \overline{y})$. This duality of notation is standard.

Was the Roman who had the job of estimating the least number of square tiles required for this mosaic the least squares estimator?

The least squares estimator of α is given by

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{x}.$$

The sampling distributions and some properties of the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ will be given in Subsection 4.1. There too you will find the estimator of σ^2, the final unknown parameter in linear regression (which we haven't addressed yet), and its sampling properties. The properties of Subsection 4.1 are used to provide a particular hypothesis test in Subsection 4.2. Other aspects of statistical inference in regression are reviewed briefly in Subsection 4.3. Note that it is not the intention of this section to provide anything like an exhaustive list of results, or to offer illustrations of many of the questions that might arise in a regression context.

4.1 The sampling distributions of the estimators

By assuming only that the random terms W_i are independent with zero mean and variance σ^2, it is a cumbersome task, although not a difficult one, to show that the two estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased estimators of α and β, respectively, that is,

$$E(\widehat{\alpha}) = \alpha, \quad E(\widehat{\beta}) = \beta.$$

Furthermore, it can be shown that the variances of the estimators are given by the formulas

$$V(\widehat{\alpha}) = \left(\frac{\overline{x}^2}{S_{xx}} + \frac{1}{n}\right)\sigma^2, \quad V(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}}.$$

The calculations involved are not entirely straightforward.

Notice that if the x values are widely dispersed, that is, if the sum of squared deviations S_{xx} is large, then the variances of both estimators are smaller than if the x values are close together (so that S_{xx} is small). This makes sense because there is more information about the parameters of the regression line when the explanatory variable takes on a wide range of values than there is if it is confined to a narrow range. As the estimators are unbiased for any values of the explanatory variable, it is possible to choose values $x_1, x_2, \ldots, x_n$ such that the variances of the estimators $\widehat{\alpha}$ and

$\widehat{\beta}$ are small, and thus the precision of the results is improved. In particular, it is helpful to obtain data for a wide spread of x values rather than to concentrate on only a narrow range. This is important when designing a statistical experiment.

Until now, the parameter σ^2 has been treated as if its value was known. In general, though, its value is not known, and it has to be replaced with an estimate. Write $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}\, x_i$ for the fitted values in random variable form. The unbiased estimator for σ^2 that is used is

$$S^2 = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2}. \tag{10}$$

Thus the numerator in S^2 is simply the residual sum of squares for the least squares line: $\sum (Y_i - \widehat{Y}_i)^2 = \sum (Y_i - (\widehat{\alpha} + \widehat{\beta}\, x_i))^2$. Convention then dictates that the unbiased estimate of σ^2 is used rather than the maximum likelihood estimate of σ^2 under the assumption of normality for the random terms W_i. As for estimation of the variance in a single sample (Unit 7), maximum likelihood estimation of the variance, σ^2, in the regression model would result in a divisor of n in Equation (10), not $n-2$. But it turns out that as unbiasedness is achieved in the one-sample case by subtracting 1 from n because there is one other parameter – the mean, μ – also being estimated, so in regression unbiasedness is achieved by subtracting 2 from n, there being two other parameters, α and β, also being estimated.

The results given so far in this subsection hold whatever form is taken by the distribution of the random terms. The results that are given in the box that follows hold when the random terms W_i are assumed to be normally distributed. This assumption is made throughout the rest of this section. The results are stated without proof.

Distributions of the least squares estimators

Assuming that, independently, $W_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$, it can be shown that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right), \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2),$$

and these two random variables are independent. These results can be combined to give the following result, which proves to be very useful for statistical inference when σ^2 is unknown (which is usually the case):

$$\frac{\widehat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t(n-2). \tag{11}$$

4.2 Testing whether a relationship exists

When $\beta = 0$, the linear regression model simplifies to

$$Y_i = \alpha + W_i.$$

In this case, the value of Y_i does not depend on the value of x_i – that is, the response variable and the explanatory variable are unrelated: it is often said that no regression relationship exists. Equivalently, the regression line is flat; see Figure 31.

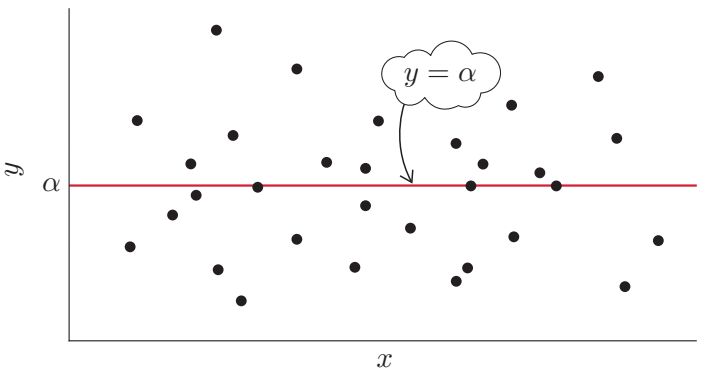


Figure 31 Artificial data from the regression $Y_i = \alpha + W_i$

In fact, if $\beta = 0$, the responses are just a random sample from a normal distribution with mean α and variance σ^2. So researchers are often interested in testing whether the slope parameter β in the linear regression model is zero, that is, testing the null hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. This can be done using Distributional Result (11).

Example 15 *Does caffeine have an effect on tapping performance?*

Tap dancers give a different type of tapping performance; do they need caffeine?

In Activity 12, you found that the equation of the least squares regression line for the finger-tapping data is

$$y = 244.75 + 0.0175\,x,$$

that is,

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose},$$

where taps are counted per minute and the caffeine dose is measured in mg.

An interesting question to consider is 'Does caffeine really have any effect on tapping frequency?' When $\beta = 0$, there is no relationship between the explanatory variable and the response variable. So one approach to answering the question is to carry out a two-sided test of the null hypothesis

$$H_0 : \beta = 0,$$

against

$$H_1 : \beta \neq 0.$$

Certainly, the estimated value $\widehat{\beta} = 0.0175$ seems to be quite a small number in absolute terms, but it needs to be assessed in the context of the overall variation in the data.

The following summary statistics are required in order to perform the test:

$$n = 30, \quad \sum (y_i - \widehat{y}_i)^2 = 134.25, \quad S_{xx} = 200\,000.$$

First, we need an estimate of σ^2. Using the sample version of Equation (10), the estimate of σ^2 is given by

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{134.25}{28} \simeq 4.7946.$$

$n - 2 = 30 - 2 = 28$

Using Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(28)$. Then, when H_0 is true, $\beta = 0$ and Distributional Result (11) means that the observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.0175 - 0}{\sqrt{4.7946}/\sqrt{200\,000}} \simeq 3.574.$$

From the table of quantiles of the t-distribution in the Handbook, the 0.999-quantile of $t(28)$ is 3.408, so the p-value for this two-sided test is less than 0.002. This p-value is extremely small, so the null hypothesis $H_0 : \beta = 0$ is rejected. That is, despite the seemingly small value of $\widehat{\beta}$, there is strong evidence against the hypothesis that caffeine dose has no effect on tapping performance.

A computer gave 0.0013 for the p-value.

Activity 17 *Does cholesterol really change with age for older ages?*

Consider again the cholesterol data for the 11 patients aged over 40 that have been much studied in previous sections. The equation of the least squares line for the cholesterol data, which was found in Example 11, is

$$y = 1.89 + 0.04\,x,$$

where y represents the total cholesterol in mg/ml, and x represents the patient's age in years. As in Example 15, the value of $\widehat{\beta} = 0.04$ seems rather close to zero, but the presence or otherwise of a non-zero slope needs to be tested taking into account the variability in the data. In order to test $H_0 : \beta = 0$ you will need the following summary statistics:

$$n = 11, \quad \sum(y_i - \widehat{y}_i)^2 = 0.952, \quad S_{xx} = 352.18.$$

(a) What is the value of s^2, the estimate of σ^2?

(b) Using a two-sided alternative hypothesis, test whether there is really no relationship between cholesterol and age.

We are not considering the larger cholesterol dataset of Exercise 2 here.

4.3 Some brief intervals

The results of Subsection 4.1 also allow us to provide interval estimators of a number of quantities associated with the linear regression model. Since these formulas are rather similar to the t-intervals of Section 4 of Unit 8, to avoid too much repetition and tedium, we do little more than list the formulas here, along with some brief comments and a single activity.

A more enjoyable sort of interval, at a cultural event

Associated with the test of $H_0 : \beta = 0$ that we have just considered in Subsection 4.2 is a confidence interval for the value of the slope parameter, β. This too arises from manipulation of Distributional Result (11). Throughout this subsection, write s for the estimated standard deviation, which is the square root of the estimated variance

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2}.$$

A $100(1-\alpha)\%$ confidence interval for the slope parameter β

A $100(1-\alpha)\%$ **confidence interval for the slope parameter** β of the regression line is given by

$$\left(\widehat{\beta} - t\,\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t\,\frac{s}{\sqrt{S_{xx}}}\right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

In a linear regression model, the mean of the response Y_i is $\alpha + \beta x_i$, that is, it depends on the value of the explanatory variable x_i. So a confidence interval for the *mean response* will also depend on the value of the explanatory variable, and therefore varies for different values of x. Suppose we are interested in the mean response *for a given value* x_0 *of* x, that is, $\alpha + \beta x_0$. The natural point estimator of $\alpha + \beta x_0$ is

$$\widehat{\alpha} + \widehat{\beta} x_0,$$

which turns out to be an unbiased estimator of $\alpha + \beta x_0$. The corresponding interval estimator of $\alpha + \beta x_0$ is given next.

A $100(1-\alpha)\%$ confidence interval for the mean response

A $100(1-\alpha)\%$ **confidence interval for the mean response** of Y_0, $\alpha + \beta x_0$, is given by

$$\left(\widehat{\alpha} + \widehat{\beta} x_0 - t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}}, \widehat{\alpha} + \widehat{\beta} x_0 + t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}}\right), \tag{12}$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

$t\,s$ is not a new symbol, just the product of the quantile, t, and the estimated standard deviation, s.

Suppose, finally, that there is interest in predicting the value of the response Y_0 for a given value x_0 of the explanatory variable. Then the obvious *predictor* of Y_0 is

$$\widehat{y}_0 = \widehat{\alpha} + \widehat{\beta} x_0.$$

Of course, you saw and used this quantity in Subsection 2.3.

This is precisely the same as the point estimator of the mean response at x_0 given above, that is, the point predictor and the point estimator of the mean response are the same. There is, however, a difference between the confidence interval for the mean response (given above) and the confidence

interval for the predicted response – the interval predictor, or *prediction interval* (given below). This is because there are *two* sources of variation in connection with prediction.

First, there is the variability associated with the least squares line that estimates the mean of the response, this variability being used in forming the confidence interval for the mean response above. In addition, though, for a given value x_0 of the explanatory variable, the response is a random variable:

$$Y_0 = \alpha + \beta x_0 + W_0.$$

So, as well as the variability associated with estimating the line at x_0, $\alpha + \beta x_0$, there is the added variation due to the random term, W_0. (Because of W_0, even if the true values of α and β were known, it would still not be possible to predict Y_0 exactly!) In a prediction interval, we have to allow for variation coming from the random term W_0 in addition to variation coming from the estimation of the predictor.

A $100(1-\alpha)\%$ prediction interval for the response

A $100(1-\alpha)\%$ **prediction interval** for the response Y_0 when $x = x_0$ is given by

$$\left(\widehat{\alpha} + \widehat{\beta} x_0 - t\, s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1},\ \widehat{\alpha} + \widehat{\beta} x_0 + t\, s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right), \tag{13}$$

where t is the $(1 - (\alpha/2))$-quantile of $t(n-2)$.

As you can see, prediction intervals are calculated in a similar way to confidence intervals. A prediction is made; and lower and upper limits are calculated, allowing for error in the prediction. However, a prediction interval has to allow for more variation than a confidence interval does. So prediction intervals are wider than confidence intervals. (There is an extra term of '1' added beneath the square root signs in Interval (13), compared with Interval (12).)

Note that the quantiles of the t-distribution required in the intervals in the boxes above are the same in all three cases.

Activity 18 *Intervals from the finger-tapping data*

In Activity 12, you fitted the following model to the finger-tapping data:

$$y = 244.75 + 0.0175\, x.$$

Suppose that the researchers were interested in the frequency of finger-tapping when an $x_0 = 40$ mg dose of caffeine was administered.

The estimated mean tapping frequency in response to a dose of $x_0 = 40$ mg of caffeine is

$$244.75 + 0.0175 \times 40 = 245.45$$

taps per minute.

The following summary statistics for these data will be needed:

$$n = 30, \quad \overline{x} = 100, \quad S_{xx} = 200\,000.$$

Also, the estimate $s^2 = 4.7946$ was found in Example 15.

(a) You will be concerned with 95% intervals, for the mean response and for the predicted response, in this question. You will require the value of t to be an appropriate quantile of an appropriate t-distribution. Find the value of t.

(b) Obtain a 95% confidence interval for the mean tapping frequency of individuals receiving a dose of $x_0 = 40$ mg of caffeine.

(c) Obtain a 95% prediction interval for the tapping frequency of a particular individual receiving a dose of $x_0 = 40$ mg of caffeine.

Hypothesis tests, confidence intervals, and so on, in relation to linear regression models can be calculated using Minitab. However, we will not spend time on this at this juncture.

Exercise on Section 4

Exercise 3 *A little inference on the full cholesterol dataset*

Consider again the cholesterol data for the full set of 24 hyperlipoproteinaemia patients, with ages from 20 years upwards, that was considered in Exercise 2. The equation of the least squares line for these data, which was found in Exercise 2, is

$$y = 1.28 + 0.05\,x,$$

where y represents the total cholesterol in mg/ml, and x represents the patient's age in years. The summary statistics needed to answer this question are as follows:

$$n = 24, \quad \overline{x} = 39.42, \quad S_{xx} = 4139.77, \quad \sum (y_i - \widehat{y}_i)^2 \simeq 2.455.$$

(a) Using a two-sided alternative hypothesis, test whether there is actually no regression relationship between age and cholesterol over the wide range of ages in the dataset.

(b) Calculate a 90% prediction interval for the value of total cholesterol for a hyperlipoproteinaemia patient of age 35 years.

(c) The prediction interval that you calculated in part (b) is actually rather wide. By reference to the data in Table 8, explain why this interval is still useful, despite its width.

5 Multiple regression

In the final section of this unit, we consider the situation in which there is more than one explanatory variable. You have already seen an example of such a scenario in Example 4 in which the response variable was the strength of timber beams and there were two possible explanatory variables: specific gravity and moisture content. In such situations, the linear regression model can be extended to incorporate more than one explanatory variable into the model; this is called **multiple regression**. Multiple regression is an important statistical method which is widely used by practising statisticians; this section provides just a brief introduction to the topic.

There continues to be a single response variable.

The section begins in Subsection 5.1 by extending the linear regression model to incorporate more than one explanatory variable into the model. The interpretation of the model parameters is not the same in multiple regression as it is in linear regression with one explanatory variable; this is discussed in Subsection 5.2. Checking the model assumptions in multiple regression is the subject of Subsection 5.3. Finally, multiple regression in Minitab is the subject of Subsection 5.4 (and its associated chapter in Computer Book C).

5.1 Extending the linear regression model

We start this subsection with an example of a problem in which there are two potential explanatory variables.

Example 16 *Student satisfaction*

Official statistics concerning UK universities are collected annually. This example considers three of the variables on which data were collected for 2015. The example focuses on data for the 24 UK universities known collectively as Russell Group universities. (This group represents some of the leading UK universities.)

The National Student Survey (NSS) surveys final-year UK undergraduate students. Surveyed students score how satisfied they are with the quality of various aspects of the teaching that they received, using a scale from 0 to 5 (where 5 represents the highest level of satisfaction). The first variable in this example is an overall student satisfaction score for each university: this is an average of the individual student satisfaction scores within that university for 2015. Although individual scores are discrete, with range $\{0, 1, 2, 3, 4, 5\}$, scores for whole universities have quite a narrow range of essentially continuous values, and for these data range from 3.89 to 4.18. Student satisfaction is our response variable Y.

As before, an explanatory variable is regarded as non-random for the purposes of regression modelling, and so is denoted by a lower-case letter.

The second variable that we'll consider here is the student–staff ratio. For each university, this is the total number of undergraduate and postgraduate students for 2015 divided by the number of academic staff for that year. The data for this variable were collected by the Higher Education Statistics Agency (HESA). Denote this explanatory variable by x_1.

The final variable that we'll consider is 'academic services spend'. These data were also collected by HESA and use the average expenditure over three academic financial years (2012/13, 2013/14 and 2014/15) to allow for uneven patterns of expenditure. Academic services spend was calculated as being the expenditure, in pounds, on library and computing facilities (staff, books, journals, computer hardware and software, but not buildings), museums, galleries and observatories, divided by the number of full-time equivalent students in the latest academic year. Denote this second explanatory variable by x_2.

The data are given in Table 9.

Table 9 Student satisfaction in Russell Group universities, 2015

Student satisfaction	Student–staff ratio	Academic services spend (£)
4.08	14.0	1883
3.96	13.8	1453
4.17	11.0	2628
4.09	13.7	1245
4.14	14.7	1542
3.93	12.0	1436
4.18	15.8	1689
4.12	14.5	1702
4.15	11.1	2309
3.91	11.7	1499
4.16	13.5	1970
4.01	11.8	1685
3.89	11.6	2105
4.05	13.4	1513
4.14	15.5	1352
4.06	13.2	1594
4.17	10.5	2700
4.12	11.9	1548
4.13	15.1	1266
4.14	14.6	1540
4.08	12.0	1694
3.95	10.2	2212
4.09	12.5	1826
4.14	14.5	1441

Figure 32 shows a scatterplot of student satisfaction (y) against the first explanatory variable, student–staff ratio (x_1). Included on the plot is the least squares line, calculated using the method given in Subsection 2.3 as

$$y = 3.797 + 0.0215\,x_1.$$

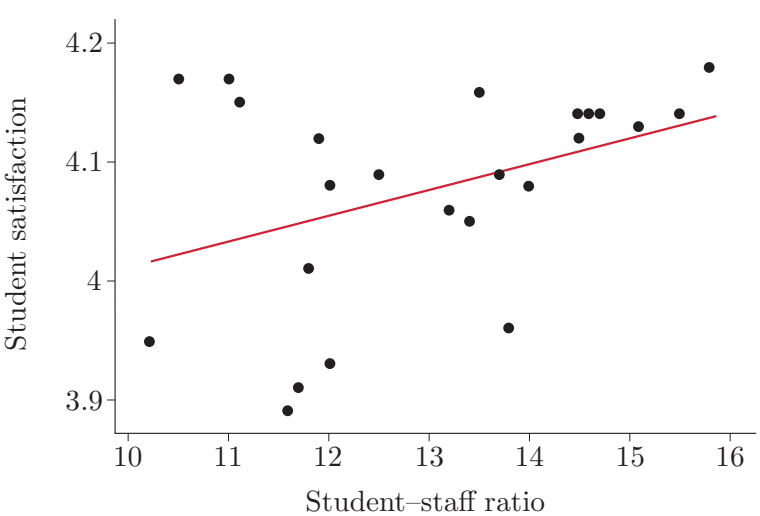


Figure 32 Student satisfaction against student–staff ratio, and the least squares line

From Figure 32, student satisfaction generally increases as the student–staff ratio increases. This is reflected in the positive slope parameter in the least squares line. You might have expected the student satisfaction to decrease as the student–staff ratio increases; and indeed this is the case when all UK universities are considered. The observed increase when considering only Russell Group universities therefore seems to be specific to these universities. (For instance, it's possible that the student–staff ratio could be a reflection of the popularity and quality of some Russell Group universities, which can attract large numbers of applicants.)

Now consider the second explanatory variable, academic services spend (x_2). Figure 33 shows a scatterplot of student satisfaction (y) against x_2, together with the least squares line

$$y = 4.019 + 0.000034\,x_2.$$

Although the relationship between this second explanatory variable and the response variable appears weak, what relationship there is appears to be positive, indicating that student satisfaction increases (slightly) as academic services spend increases.

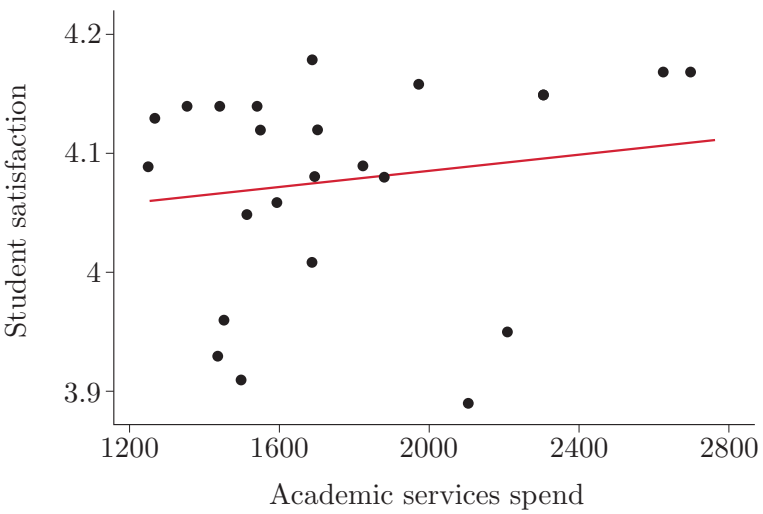


Figure 33 Student satisfaction against academic services spend, and the least squares line

When carrying out a two-sided test of the null hypothesis $H_0 : \beta = 0$ in the regression model using x_1, the associated p-value for the slope is 0.057, and when carrying out the same test in the regression model using x_2, the p-value for the slope is 0.486. So, in actual fact, there is only weak evidence that student–staff ratio on its own affects student satisfaction, and there is little or no evidence that academic services spend on its own affects student satisfaction. Is it possible, however, that student–staff ratio and academic services spend can act *together* to affect student satisfaction in Russell Group universities in a rather more substantial way? You will see that, by using a regression model which uses *both* explanatory variables at the same time, this is indeed the case!

These p-values were obtained from Minitab.

In Example 16, we had a response variable Y with two explanatory variables x_1 and x_2. Denote the ith observations of x_1 and x_2 (associated with y_i) by x_{i1} and x_{i2}, respectively. Now, the linear regression model with one explanatory variable can be written as

$$Y_i = \alpha + \beta\,x_i + W_i,$$

where the W_is are independent random variables with zero mean and constant variance. This model can be extended to incorporate two

explanatory variables thus:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i.$$

Once again the W_is are independent random variables with zero mean and constant variance. In fact, here we will consider only the case in which the W_is are additionally assumed to be normally distributed.

This model can be naturally extended to the situation in which there are p explanatory variables $x_1, x_2, \ldots, x_p$, with the ith observation of the jth explanatory variable being denoted by x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. The *multiple linear regression model*, or the *multiple regression model* for short, is then defined as follows.

The multiple linear regression model

If data $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, $i = 1, 2, \ldots, n$, comprise observations on p explanatory variables $x_1, x_2, \ldots, x_p$ and a response variable Y, then the **multiple linear regression model** can be written

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + W_i, \qquad (14)$$

$i = 1, 2, \ldots, n$. The terms W_i are independent normal random variables with zero mean and constant variance.

Note that we are considering only the situation in which the relationship between Y and $x_1, x_2, \ldots, x_p$ is linear and the random terms W_i, $i = 1, 2, \ldots, n$, come from independent normal distributions.

Activity 19 *Formulating a model*

A zoologist would like to use a multiple linear regression model to model the heights of young giraffes using their weight and age (in days). Write down the form of the zoologist's multiple regression model.

5.2 Interpreting regression coefficients

So, how are the parameters of the multiple linear regression given in Equation (14) to be interpreted?

Well, first, the parameter α can still be considered as an intercept parameter because it is the value of the linear trend $\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ when all of $x_{i1}, x_{i2}, \ldots, x_{ip}$ are zero.

The parameters $\beta_1, \beta_2, \ldots, \beta_p$, however, are now **partial regression coefficients**. They are usually just called **regression coefficients** for short, but the word 'partial' is important in reminding us of their meaning. In the multiple regression model, the parameter β_1 measures the effect of increasing x_1 by one unit when $x_2, x_3, \ldots, x_p$ *are all kept fixed*; β_2 measures the effect of increasing x_2 by one unit when $x_1, x_3, \ldots, x_p$ *are all kept fixed*;

and so on. As such, the regression coefficients are not the same as the slope parameter in the simple linear regression model with one explanatory variable, and they do not have the same interpretation: a regression coefficient represents the 'partial' effect of the associated explanatory variable given the values of the other explanatory variables, while the slope parameter represents the effect of a single explanatory variable on its own.

In Section 2, the method of least squares was used to estimate the parameters in the linear regression model with a single explanatory variable. You also saw, in Subsection 2.4, that when the random terms W_i are normally distributed, maximum likelihood estimates of the parameters of the linear regression model are the same as those obtained via the method of least squares. Parameter estimation when there is more than one explanatory variable follows the same ideas, but is a bit more complicated due to the increased number of parameters. Because of this, in M248 we will simply use Minitab for estimating the intercept parameter and regression coefficients. (Details of estimation are left to modules at a higher level.)

The fitted multiple regression model

If $\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ are estimates of the intercept and regression coefficients in a multiple regression model, then the fitted multiple regression model is

$$y = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p.$$

Example 17 *Interpreting the fitted model*

Consider once again Example 16 in which we had the response variable student satisfaction (Y) and two explanatory variables: student–staff ratio (x_1) and academic services spend (x_2). The fitted multiple regression model obtained by using Minitab with the data of Table 9 is

$$y = 3.157 + 0.0484\, x_1 + 0.000166\, x_2.$$

The interpretation of the regression coefficients is as follows.

- If the value of the student–staff ratio (x_1) increases by one unit (that is, by one more student per staff member), and the value of the academic services spend (x_2) remains fixed, then the student satisfaction score (y) would be expected to increase by 0.0484.
- If the value of the academic services spend (x_2) increases by one unit (that is, by one pound per student), and the value of the student–staff ratio (x_1) remains fixed, then the student satisfaction score (y) would be expected to increase by 0.000166.

Notice that the regression coefficients are not the same values as the corresponding slope parameters for x_1 and x_2 in the separate least squares lines in Example 16. In each least squares line, the slope parameter represents the effect of the individual explanatory variable on the response

variable. However, when both explanatory variables are in the model, as they are here, each regression coefficient represents the partial effect that the individual explanatory variable has on the response variable, given the other explanatory variable.

In Example 16, you saw that there wasn't very much evidence to suggest that either of the slope parameters in the separate linear regression models for modelling student satisfaction was non-zero. So, treated individually, it looked likely that neither the first explanatory variable, student–staff ratio, nor the second explanatory variable, academic services spend, was going to be very useful for modelling student satisfaction. How do we know that the regression model with both explanatory variables given in Example 17 is any better? The answer lies in carrying out a two-sided test of the null hypothesis

$$H_0 : \beta_1 = 0,$$

and a second two-sided test of the null hypothesis

$$H_0 : \beta_2 = 0,$$

within the context of the multiple linear regression model with two explanatory variables. These tests are similar in construction to the two-sided test in the linear regression model with one explanatory variable of the null hypothesis $H_0 : \beta = 0$. But the pair of multiple regression tests yields different results from the pair of simple linear regression tests, for reasons again associated with the partial nature of the regression coefficients in the multiple regression context. You will be spared the details of these tests here, and instead we will just consider the resulting p-values which are routinely provided by Minitab when fitting a multiple regression model.

These p-values are used to assess the evidence against the null hypothesis in the usual way.

Activity 20 *Are the student satisfaction regression coefficients zero?*

The fitted multiple regression model for response variable student satisfaction (Y) and two explanatory variables, student–staff ratio (x_1) and academic services spend (x_2), obtained from Minitab is

$$y = 3.157 + 0.0484\,x_1 + 0.000166\,x_2.$$

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is calculated in Minitab to be 0.000, and the p-value for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is calculated in Minitab to be 0.002. What do you conclude about the regression coefficients for x_1 and x_2? Hence, what do you conclude about how student–staff ratio and academic services spend affect student satisfaction in Russell Group universities?

As was the case with a single explanatory variable, the fitted multiple linear regression line can be used for prediction. Point prediction is particularly straightforward and is illustrated in the context of student satisfaction scores in Example 18.

Example 18 *Predicting student satisfaction*

Suppose that another university felt that the Russell Group fitted line applied equally well to it. In 2015, this university had a student–staff ratio of 14.5 students per staff member and an academic services spend of £1441 per student. The fitted line predicts a student satisfaction score of

$$3.157 + 0.0484 \times 14.5 + 0.000166 \times 1441 \simeq 4.10.$$

(Perhaps this university was right: its actual 2015 student satisfaction score turns out to have been 4.14.)

Activity 21 *Strength of timber beams*

Example 4 introduced a dataset involving timber beams. The response variable Y is the strength of a timber beam, and there are two explanatory variables, specific gravity (x_1) and moisture content (x_2). Scatterplots of y against x_1 and of y against x_2 were given in Figure 5. These suggested that there may be an increasing linear relationship between y and x_1, but a weaker, decreasing, relationship between y and x_2. We can use multiple regression to investigate whether specific gravity and moisture content together affect the strength of timber beams.

The fitted multiple regression model for this dataset, obtained from Minitab, is

$$y = 10.29 + 8.50\,x_1 - 0.265\,x_2.$$

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is 0.002, while that for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is 0.069.

(a) Interpret the regression coefficients.

(b) Do the data suggest that both x_1 and x_2 together influence the strength of timber beams?

(c) Using the fitted multiple regression model, predict the strength of a timber beam with specific gravity 0.5 and moisture content 10.

In the next activity you will consider a dataset in which there are more than two explanatory variables.

Activity 22 *Gross domestic product*

The average level of income per person varies widely across countries and changes over time as some countries decline and others grow. Economists are interested in the question: 'Why do some countries grow faster than others?' In this activity, a multiple regression model is used to investigate this question. Economic data for 128 countries are available. The response

variable, Y, is the rate of growth, specifically the rate of change between 2000 and 2010 of the gross domestic product (GDP) per head, where the GDP is the total output produced in the country in one year per person. In the dataset, the growth is given as a decimal rather than as a percentage.

There are three explanatory variables, x_1, x_2 and x_3.

There will be consideration of transformations of variables within regression models in Unit 12; just take this logarithmic transformation as providing a helpful scale for this variable in this case.

- x_1 is a measure of the output (GDP) per head in 2000, the initial year of the period; more specifically, it happens to be the logarithm of the GDP per head, where GDP has been translated to the value in US dollars from 2005. Differences in GDP per head are related to differences in the level of technology used. Countries that are more technologically advanced tend to have high levels of GDP per head, while countries that tend to use older and less efficient technology tend to have low levels of GDP per head. Since it is much more difficult and expensive to generate technological innovation than to copy existing technology, it should be easier for poorer countries to grow faster than richer countries by copying better existing technology and therefore improving their efficiency. In turn, this means that countries with low initial GDP per head in 2000 have greater scope for growing and therefore catching up with richer countries.

- x_2 is the share of gross fixed capital formation in GDP in the ten-year period. This is a percentage. Gross fixed capital formation is the investment in new plants, machinery and equipment that is necessary to produce more output (goods and services) and is considered by economists to be a key engine of growth. Intuitively, the argument runs as follows. The output produced can consist of either consumer goods which are used up, such as a loaf of bread, or capital goods which are used as inputs to produce new output in the future, such as a new milling machine. Countries that invest more by producing a greater share of capital goods increase their stock of capital available for production of future output, so they should grow faster than countries that focus more on consumption. So a high share of gross fixed capital formation in GDP should be associated with higher growth.
- x_3 is the total enrolment in secondary schools. This too is a percentage, in this case of the population aged 15 or over. The total enrolment in secondary schools is a measure of human capital, the level of education of the workforce, which is associated with higher productivity and therefore faster economic growth.

The fitted multiple regression model for this dataset, obtained from Minitab, turns out to be

$$y = 0.312 - 0.0923\,x_1 + 0.02425\,x_2 + 0.00493\,x_3.$$

The p-value for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, is reported by Minitab to be 0.000.

(a) Explain why this analysis suggests that all three explanatory variables together influence the rate of growth of GDP.

(b) Interpret each of the regression coefficients.

(c) Predict what the rate of growth between 2000 and 2010 would have been for a fictional South Asian country whose 'logged' output per head in 2000 (in the appropriate units) was 6, whose gross fixed capital formation share was 25%, and whose total enrolment in secondary schools was 40%.

5.3 Checking the assumptions

An essential part of any regression analysis is to check the model assumptions. For the multiple linear regression model we have the following assumptions.

1 The random terms W_i are independent.

2 The W_is have zero mean and constant variance.

3 The W_is are normally distributed.

You might be thinking that these three assumptions look very familiar, and you would be right! We have exactly the same assumptions for the multiple linear regression model that we had for the linear regression model with one explanatory variable. As such, these assumptions can be checked in exactly the same way. (As for simple linear regression, although Assumption 1, the independence of the W_is, can be checked, we will not do so in M248.)

In much the same way as for linear regression with one explanatory variable, the fitted multiple regression model can be used to calculate the

fitted value of the response, $\widehat{y}_i$, given values of the explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$. The formula is

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}.$$

The fitted values can then be used to calculate the residuals

$$w_i = y_i - \widehat{y}_i.$$

This is illustrated in the next example.

Example 19 *Calculating a fitted value and residual*

From Example 17, the fitted multiple regression model for response variable student satisfaction (Y) and two explanatory variables, student–staff ratio (x_1) and academic services spend (x_2), obtained from Minitab is

$$y = 3.157 + 0.0484\, x_1 + 0.000166\, x_2.$$

A very predictable university?!

Out of the 24 Russell Group universities in the sample, the University of Liverpool achieved a student satisfaction score of 4.01. For this university, the student–staff ratio was 11.8, while the academic services spend was £1685 per student. The University of Liverpool is the 12th university listed in the sample. The fitted value of student satisfaction for Liverpool is then

$$\widehat{y}_{12} = 3.157 + 0.0484 \times 11.8 + 0.000166 \times 1685 = 4.0078 \simeq 4.01.$$

The associated residual is therefore

$$w_{12} = 4.01 - 4.01 = 0.$$

So, for this university, the actual student satisfaction score is the same value as the fitted student satisfaction score estimated from the multiple regression model. The values of student–staff ratio and academic services spend allow us to predict the student satisfaction score very well.

Activity 23 *More fitted values and residuals*

For the University of Exeter, the student satisfaction score was 4.18, the student–staff ratio was 15.8, and the academic services spend was £1689 per student. The University of Exeter is the 7th university listed in the sample. For Queen Mary University of London, the student satisfaction score was 4.12, the student–staff ratio was 11.9, while the academic services spend was £1548 per student. Queen Mary is the 18th university listed in the sample.

Calculate the residuals w_7 and w_{18}. Comment on the values of the residuals you obtain.

With fitted values and residuals in place, you should be able to do Activity 24.

Activity 24 *How to check the model assumptions*

(a) Explain how you would check that Assumption 2, that the W_is have zero mean and constant variance, is reasonable.

(b) Explain how you would check that Assumption 3, that the W_is are normally distributed, is reasonable.

You will check the model assumptions for the university student satisfaction data in the following activity.

Activity 25 *Checking the assumptions for the student satisfaction model*

Figure 34 shows the residual plot and the normal probability plot of the residuals for the fitted multiple linear regression model given in Example 17 for the university student satisfaction dataset first considered in Example 16.

Do these plots suggest that the model assumptions are reasonable?

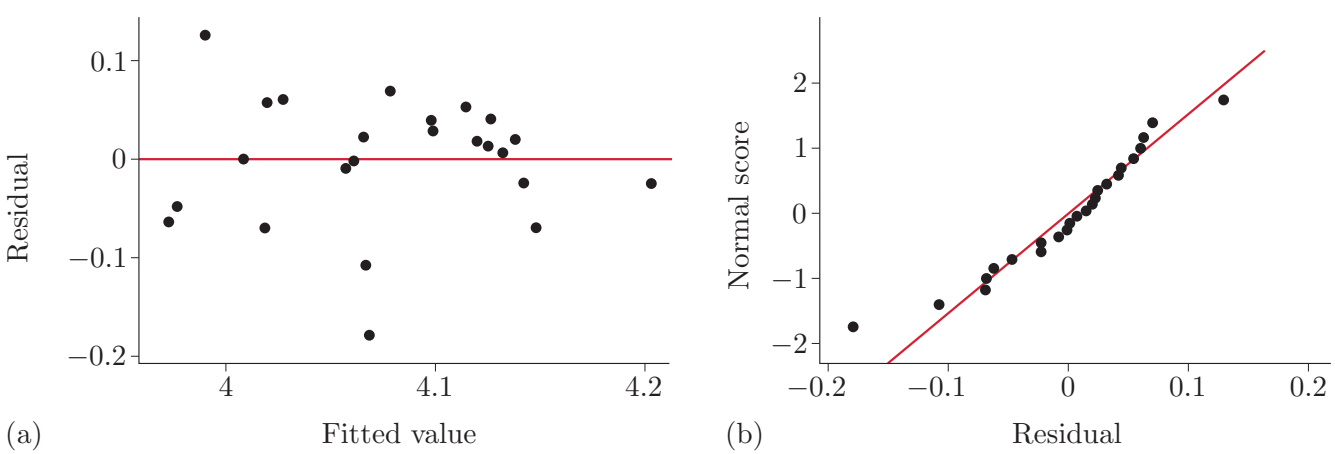


Figure 34 Checking the assumptions for the student satisfaction data: (a) residual plot; (b) normal probability plot

5.4 Multiple regression in Minitab

The final part of this section involves using multiple linear regression in Minitab.

Refer to Chapter 3 of Computer Book C for the rest of the work in this section.

Exercise on Section 5

Exercise 4 *Another multiple regression model for growth of GDP*

'Prevalence' is a word often used for a proportion or percentage when talking about medical conditions.

In Activity 22, a multiple regression model was fitted to data from 128 countries with response variable the rate of growth of gross domestic product (Y) over 2000–2010, and three explanatory variables: log of output per head in 2000 (x_1), share of gross fixed capital formation in the ten-year period (x_2), and total enrolment in secondary school (x_3). There is also available a fourth explanatory variable, x_4, the prevalence of HIV as a proportion of population for ages 15–49. The prevalence of HIV is a factor that might affect growth in some poorer countries because a high prevalence of HIV can reduce the contribution of productive workers. Data for this explanatory variable were available for only 78 of the 128 countries considered in Activity 22. A multiple regression model with all four explanatory variables was fitted for the 78 countries with complete data. The fitted model is

$$Y = 0.357 - 0.0895\,x_1 + 0.02118\,x_2 + 0.00519\,x_3 - 0.00892\,x_4.$$

The p-values for individual two-sided tests of the hypotheses $H_0 : \beta_j = 0$, are 0.000 for $j = 1, 2, 3$ and 0.032 for $j = 4$. The residual plot and normal probability plot of residuals are given in Figure 35.

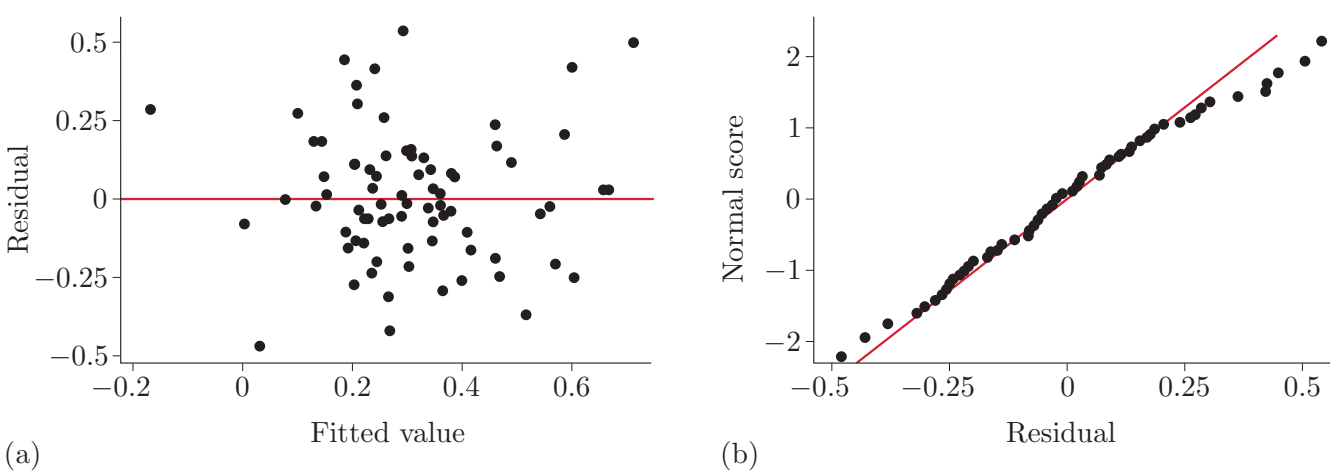


Figure 35 Checking the assumptions for the GDP dataset: (a) residual plot; (b) normal probability plot

(a) Explain why this analysis suggests that all four explanatory variables together affect the rate of growth of GDP.

(b) Do the model assumptions seem reasonable?

(c) Interpret the regression coefficients in the fitted model.

(d) For the fictional South Asian country in Activity 22(c) whose log of output per head in 2000 was 6, whose gross fixed capital formation share was 25%, whose total enrolment in secondary schools was 40%, and whose HIV prevalence is 0.1, use the fitted multiple regression model to predict its growth in GDP. How does this prediction compare

with the prediction you made on the basis of the multiple regression model with just three explanatory variables in Activity 22(c)?

Summary

In this unit, you have learned about regression models. The general regression model has been defined and a particular simple case, the linear regression model with one explanatory variable, has been treated in some depth. You have learned how to fit linear regression models to data, and to check the assumptions of a fitted model. Also, you have learned how to test whether there really is any linear regression relationship at all, and have briefly explored confidence intervals for the slope and for the mean response for a given value of the explanatory variable, and prediction intervals for the response for a given value of the explanatory variable. Finally, you saw how linear regression can be extended to incorporate more than one explanatory variable through multiple regression.

You have used Minitab to fit linear regression models, both simple and multiple, and to produce appropriate plots in order to check the modelling assumptions.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate that an explanatory variable might be thought of as 'explaining' the value of another (response) variable, and that a response variable 'responds' to the value of one or more other (explanatory) variables
- understand that a general regression model contains a function describing how the response variable is related to the explanatory variable, and a random term which models the variation in the response
- appreciate that a linear regression model is a special case of the general regression model in which the relationship between the variables is linear
- understand that the random terms in the linear regression model are assumed to be independent with constant, zero mean and constant variance
- use a scatterplot to decide if a regression model (or a linear regression model) might be an appropriate model for the data

- fit a straight-line model to data using the method of least squares, both by hand given summary statistics for the data, and using Minitab
- calculate fitted values, residuals and predicted values
- use Minitab to produce residual plots and normal probability plots of the residuals in order to check the assumptions of a linear regression model
- appreciate that if a residual plot shows a pattern, then the assumption of constant, zero mean and constant variance of the random terms might not be justified
- appreciate that if the residuals in a normal probability plot do not fall close to a straight line, then the random terms of a linear regression model might not be normally distributed
- given summary statistics for the data, test if the response variable is related to the explanatory variable in a simple linear regression model
- given summary statistics for the data, obtain a confidence interval for the mean response in a linear regression model
- given summary statistics for the data, calculate a prediction interval for the response in a linear regression model
- appreciate how the linear regression variable with one explanatory variable is extended to the multiple regression model with several explanatory variables
- interpret the (partial) regression coefficients of the multiple linear regression model
- appreciate that the assumptions of the multiple regression model are the same as those of the simple linear regression model, and use residual plots and normal probability plots of residuals in the same way to check the assumptions
- use Minitab to fit a multiple regression model.

Solutions to activities

Solution to Activity 1

It would be natural to regard height as the response variable and age as the explanatory variable. This is because age 'explains' height and it wouldn't make sense to think of height 'changing' age.

Solution to Activity 2

(a) The natural background for this example would be a paper manufacturer wishing to estimate the optimal amount of hardwood to use in production to ensure the strongest possible paper. To do this, he must know how tensile strength depends on the percentage of hardwood in the pulp. That is, tensile strength is the response variable and hardwood content is the explanatory variable.

(b) In the scatterplot in Figure 7, there is a very evident relationship between the two variables. However, the relationship is not linear. It appears (from this experiment) that kraft paper is at its strongest for some intermediate level of pulp hardwood content (about 10%). A curve (quadratic or cubic) might be useful to model the relationship.

Solution to Activity 3

An appropriate regression model for these data might also be of the form

$$Y_i = \alpha + \beta x_i + W_i.$$

As in Example 6, α and β are the intercept and slope, respectively, of the straight line relating the variables, and the W_is are random terms accounting for the scatter around the straight line. In this case, the random terms W_i might have normal distributions with zero mean and some constant variance σ^2. Moreover, the W_is are independent because the height of one schoolboy has no affect on the height of another schoolboy.

Solution to Activity 4

Since $\alpha + \beta x_i$ is a constant, use the results from Unit 4 that, for any random variable X, $E(a + bX) = a + b\,E(X)$, $V(a + bX) = b^2\,V(X)$, with $a = \alpha + \beta x_i$, $b = 1$ and $X = W_i$, to find that

$$\begin{aligned} E(Y_i) &= E(\alpha + \beta x_i + W_i) = \alpha + \beta x_i + E(W_i) \\ &= \alpha + \beta x_i + 0 = \alpha + \beta x_i, \end{aligned}$$

$$V(Y_i) = V(\alpha + \beta x_i + W_i) = V(W_i) = \sigma^2.$$

Solution to Activity 5

(a) The problem with using the sum of residuals is that positive and negative residuals (which might be quite large in absolute value) cancel each other out. By summing the squared residuals, residuals that are large in absolute value add substantially to the sum, whether they be positive or negative.

(b) Instead of summing squared residuals, you could sum the absolute values of the residuals, also forcing large residuals to contribute substantially to the sum whether they are positive or negative. Other possibilities include taking the residuals to the fourth power prior to summing.

As an aside, minimising the sum of absolute values of residuals is also quite a popular method in statistics. An advantage it has over least squares is that it is less readily influenced by outliers; a disadvantage is that it does not afford explicit formulas for parameter estimates (the sum of absolute residuals has to be minimised numerically using a computer). This method will not be considered further in this module.

Solution to Activity 6

Equation (2) gives

$$\begin{aligned} R(\gamma) &= \sum_{i=1}^{n}(y_i - \gamma x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2\gamma y_i x_i + \gamma^2 x_i^2) \\ &= \sum_{i=1}^{n} y_i^2 - 2\gamma \sum_{i=1}^{n} y_i x_i + \gamma^2 \sum_{i=1}^{n} x_i^2. \end{aligned}$$

This is of the form $a\gamma^2 + b\gamma + c$ with

$$a = \sum_{i=1}^{n} x_i^2, \quad b = -2\sum_{i=1}^{n} x_i y_i, \quad c = \sum_{i=1}^{n} y_i^2.$$

(Here, the standard convention of writing $\sum_{i=1}^{n} x_i y_i$ rather than the equivalent $\sum_{i=1}^{n} y_i x_i$ has been followed.)

Solution to Activity 7

(a) (i) Expanding the square in the right-hand side of Equation (3) and manipulating further, we find that

$$\begin{aligned} a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c &= a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) - \frac{b^2}{4a} + c \\ &= ax^2 + bx + \frac{b^2}{4a} - \frac{b^2}{4a} + c \\ &= ax^2 + bx + c, \end{aligned}$$

as required.

(ii) When $a > 0$, the expression on the right-hand side of Equation (3) comprises a positive constant times a squared term depending on x, plus constants. It is therefore minimised if we can choose x to make the squared term zero. This happens if

$$x + \frac{b}{2a} = 0,$$

that is,

$$x = -\frac{b}{2a},$$

as required.

(b) With $x = \gamma$ and, from the solution to Activity 6, $b = -2\sum_{i=1}^{n} x_i y_i$ and $a = \sum_{i=1}^{n} x_i^2$, the minimiser of $R(\gamma)$ is given by

$$\gamma = -\frac{-2\sum_{i=1}^{n} x_i y_i}{2\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Solution to Activity 8

The least squares estimate of the slope γ is

$$\widehat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{219\,817}{796\,253} \simeq 0.276.$$

The least squares line through the scattered data points has equation

$$y = 0.276\,x.$$

That is, the regression relationship between the explanatory variable and the response variable can be written

$$\text{beetle count} = 0.276 \times \text{bracket weight}.$$

(In practice, you should always obtain a scatterplot before fitting a regression model. In fact, in this case, a scatterplot suggests that an unconstrained line would be more appropriate than a line through the origin.)

Solution to Activity 9

(a) Starting from Equation (4),

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 = \sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2 = \sum x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 \\ &= \sum x_i^2 - n\overline{x}^2, \end{aligned}$$

which is the second version of Equation (7).

(b) Mathematically, the only difference is a notational change, from xs in part (a) to ys here.

(c) Starting from Equation (6),

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\,\overline{y}) \\ &= \sum x_i y_i - \overline{y}\sum x_i - \overline{x}\sum y_i + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - n\overline{y}\,\overline{x} - n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - n\overline{x}\,\overline{y}, \end{aligned}$$

which is the second version of Equation (9).

Solution to Activity 10

Using Equations (7) to (9), S_{xx}, S_{yy} and S_{xy} can be calculated from the summary statistics as

$$S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 30\,409 - \frac{575^2}{11} \simeq 352.182,$$

$$S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 179.14 - \frac{44.2^2}{11} \simeq 1.536,$$

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 2324.8 - \frac{575 \times 44.2}{11} \simeq 14.345.$$

Solution to Activity 11

When $x = \overline{x}$ is inserted in the equation for the least squares line $y = \overline{y} + \widehat{\beta}(x - \overline{x})$, we find that

$$y = \overline{y} + \widehat{\beta}(\overline{x} - \overline{x}) = \overline{y},$$

as required.

Solution to Activity 12

(a) For these data,

$$S_{xx} = 500\,000 - \frac{3000^2}{30} = 200\,000,$$

$$S_{xy} = 743\,000 - \frac{3000 \times 7395}{30} = 3500.$$

(b) The least squares estimate of the slope is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{3500}{200\,000} = 0.0175.$$

The estimate of the intercept term is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{7395}{30} - 0.0175 \times \frac{3000}{30} = 244.75.$$

The equation of the least squares line is

$$y = 244.75 + 0.0175\,x,$$

or, equivalently,

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose},$$

where taps are counted per minute, and the caffeine dose is measured in mg.

(c) The value $\widehat{\alpha} = 244.75$ is the estimated value of the intercept, that is, the value of the regression line when $x = 0$. It is meaningful in this case as the estimated value of the average number of taps per minute (or the predicted number of taps per minute) for a student in receipt of no caffeine. (As an aside, this value is not the same as the average response of the 10 no-caffeine students who happened to be measured in the experiment; it is, however, extremely close since that average happens to be 244.8.)

The value $\widehat{\beta} = 0.0175$ is the estimated value of the slope. It estimates that for each additional milligram of caffeine, a student might on average be able to increase his average number of taps per minute by 0.0175.

(d) If the caffeine dose is 50 mg, the predicted number of taps per minute is

$$244.75 + 0.0175 \times 50 = 245.625.$$

Solution to Activity 13

There is a definite pattern in the residual plot in Figure 23. The residuals are increasing at first, then there is a single large negative residual, and finally the residuals return to a high positive level before decreasing. That is, Assumption 2, that the residuals come from distributions with constant, zero mean and constant variance, appears to be violated. It seems that a linear regression model is not a good model for these data after all.

(The residual plot suggests a systematic discrepancy from linearity throughout the range of the data as well as an outlier. This is despite the claim in Example 3, based on Figure 4, that 'there may well be a straight-line relationship' and the fitting of a linear regression model in Exercise 1. As well as the outlier, which can be seen in Figure 4, perhaps the data deserve to be modelled by lines of different slope either side of the outlier.)

Solution to Activity 14

The pattern of points in the residual plot gives no reason to doubt the assumption of constant, zero mean but gives plenty of reason to doubt the assumption of constant variance. Instead, the variability of the residuals appears to increase as the sizes of the fitted values increase. (The 'band' of points widens towards the right.) The linear regression model appears not to be appropriate for these data in the sense that constant variance cannot be assumed.

(Actually, in Example 26 of Unit 1 it was commented that 'an extra feature that you might perceive in [the scatterplot of these data] is that the amount of spread of the points about any central line appears to increase as the values of the measurements increase'.)

Solution to Activity 15

$$\begin{aligned}\sum_{i=1}^{n} w_i &= \sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}\,x_i) = \sum_{i=1}^{n}\{y_i - (\overline{y} - \widehat{\beta}\,\overline{x}) - \widehat{\beta}\,x_i\} \\ &= \sum_{i=1}^{n}\{y_i - \overline{y} - \widehat{\beta}\,(x_i - \overline{x})\} = \sum_{i=1}^{n}(y_i - \overline{y}) - \widehat{\beta}\sum_{i=1}^{n}(x_i - \overline{x}) \\ &= n\overline{y} - n\overline{y} - \widehat{\beta}(n\overline{x} - n\overline{x}) = 0.\end{aligned}$$

Solution to Activity 16

There is no particular pattern in the residual plot in Figure 27(a) (other than that due to the very discrete nature of the values of the explanatory variable). It seems that Assumption 2, that the W_is come from distributions with constant, zero mean and constant variance, is a reasonable one.

Also, the normal probability plot of the residuals in Figure 27(b) appears to accommodate Assumption 3, that the W_is are normally distributed. This is because the points in the plot roughly follow a straight line; the main departures from this, if any, are due to the 'stacking up' (that is,

jittering) of points with the same value of their explanatory variables and their response variables, and hence their residuals.

Overall, the linear regression model with normally distributed random terms appears to be a reasonable one to explain the dependence of the number of taps per minute on caffeine dose.

Solution to Activity 17

(a) $s^2 = \dfrac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \dfrac{0.952}{9} \simeq 0.1058.$

(b) The null hypothesis is $H_0 : \beta = 0$. From Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(9)$. The observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.04}{\sqrt{0.1058}/\sqrt{352.18}} \simeq 2.308.$$

The 0.975-quantile of $t(9)$ is 2.262 and the 0.99-quantile of $t(9)$ is 2.821, so the p-value for a two-sided test is slightly less than 0.05. (A computer gave 0.047 for the p-value.) There is therefore moderate evidence against H_0, that there is no relationship between cholesterol and age, but with a p-value close to 0.05, the evidence is somewhat marginally moderate to weak in this case.

Solution to Activity 18

(a) To calculate 95% intervals for the finger-tapping data, the 0.975-quantile of $t(28)$ is required: from the table in the Handbook, this is $t = 2.048$.

(b) Using Interval (12), a 95% confidence interval for the mean $\alpha + 40\beta$ is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\, x_0 \pm t\, s \sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}\right)$$

$$= \left(245.45 \pm 2.048\sqrt{4.7946}\sqrt{\frac{(40-100)^2}{200\,000} + \frac{1}{30}}\right)$$

$$\simeq (245.45 \pm 1.016) \simeq (244.43, 246.47).$$

(c) Using Interval (13), a 95% prediction interval for the finger-tapping frequency attained by an individual after a 40 mg dose of caffeine is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\, x_0 \pm t\, s \sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right)$$

$$= \left(245.45 \pm 2.048\sqrt{4.7946}\sqrt{\frac{(40-100)^2}{200\,000} + \frac{1}{30} + 1}\right)$$

$$\simeq (245.45 \pm 4.598) \simeq (240.85, 250.05).$$

Notice that this prediction interval is wider than the confidence interval calculated in part (b).

Solution to Activity 19

The zoologist is interested in modelling height using the weight and age of the giraffes. So height is the response variable Y, while weight and age are the explanatory variables; let weight be denoted x_1 and age be denoted x_2. Then the multiple regression model for data on weight, age and height is

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i,$$

where W_i is a normally distributed random variable with zero mean and constant variance.

Solution to Activity 20

Since the p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$ is 0.000, this means that $p < 0.01$. Therefore there is strong evidence to suggest that β_1 is not 0, that is, the regression coefficient for x_1 is not equal to 0.

The p-value for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$ is 0.002. So once again $p < 0.01$ and there is strong evidence to suggest that β_2 is not 0, that is, the regression coefficient for x_2 is also not equal to 0.

Notice that when x_1 and x_2 were considered individually in separate linear regression models in Example 16, the p-values for the slope parameters suggested that there wasn't enough evidence to suggest that either of them was non-zero. (This was especially true for x_2.) However, when we have both x_1 and x_2 in the model, the p-values suggest that there is strong evidence that the regression coefficients for both explanatory variables are non-zero. So it looks like student–staff ratio and academic services spend work *together* to affect student satisfaction.

Solution to Activity 21

(a) The interpretation of the regression coefficients is as follows.

- If the value of specific gravity (x_1) increases by one unit, and the value of moisture content (x_2) remains fixed, then the strength of timber beams (y) would be expected to increase by 8.50.
- If the value of moisture content (x_2) increases by one unit, and the value of specific gravity (x_1) remains fixed, then the strength of timber beams (y) would be expected to decrease by 0.265. The decrease is because of the negative regression coefficient.

(b) For the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$, since $p = 0.002 < 0.01$, there is strong evidence to suggest that β_1 is not zero. However, for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$, since $p = 0.069$ satisfies $0.05 < p < 0.1$, there is only weak evidence to suggest that β_2 is not equal to zero. Therefore, overall there is only weak evidence to suggest that both x_1 and x_2 *together* influence the strength of timber beams.

(c) Using the fitted multiple regression line, a beam with $x_1 = 0.5$ and $x_2 = 10$ is predicted to have strength

$$10.29 + 8.50 \times 0.5 - 0.265 \times 10 = 11.89.$$

Solution to Activity 22

(a) The p-values for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, are 0.000, which means that for each regression coefficient $p < 0.01$. There is therefore strong evidence that each regression coefficient is non-zero, which in turn implies that together the three explanatory variables influence Y, the rate of growth of GDP.

(b) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of x_1 increases by one unit, and the values of x_2 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.0923 (or a little over 9% over the ten-year period). The decrease is because of the negative regression coefficient. From the information in the question, this makes sense because it means that poorer countries tend to catch up with richer countries by copying existing technology available on global markets, and countries who are initially richer, with higher values of x_1, will grow more slowly.
- Regression coefficient for x_2: If the value of x_2 increases by one unit, and the values of x_1 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.02425 (or about 2.4% over the ten-year period). The increase is because of the positive regression coefficient. From the information in the question, this makes sense because it suggests that countries that invest a greater share of their resources in capital goods, such as industrial plants, machinery and equipment, than consumption (and so have a higher value of x_2), grow faster than countries that focus more on consumption (and so have a lower value of x_2).
- Regression coefficient for x_3: If the value of x_3 increases by one unit, and the values of x_1 and x_2 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.00493 (or about 0.5% over the ten-year period). The increase is because of the positive regression coefficient. From the information in the question, this makes sense because an increase in enrolment in secondary school (x_3) increases the education of the workforce, which would be associated with faster economic growth and increased change in GDP.

(c) Using the fitted multiple regression line, a country with $x_1 = 6$, $x_2 = 25$ and $x_3 = 40$ is predicted to have had a growth rate over the ten-year period of

$$0.312 - 0.0923 \times 6 + 0.02425 \times 25 + 0.00493 \times 40 \simeq 0.56$$

(or about 56%).

Solution to Activity 23

The fitted student satisfaction score for Exeter is

$$\widehat{y}_7 = 3.157 + 0.0484 \times 15.8 + 0.000166 \times 1689 \simeq 4.2021 \simeq 4.20.$$

The associated residual is therefore

$$w_7 = y_7 - \widehat{y}_7 = 4.18 - 4.20 = -0.02.$$

For the University of Exeter, the student satisfaction score seems to be fairly close to the fitted student satisfaction score estimated from the multiple regression model. The values of student–staff ratio and academic services spend allow us to predict the student satisfaction score well.

The fitted student satisfaction score for Queen Mary University of London is

$$\widehat{y}_{18} = 3.157 + 0.0484 \times 11.9 + 0.000166 \times 1548 \simeq 3.9900 \simeq 3.99.$$

The associated residual is therefore

$$w_{18} = y_{18} - \widehat{y}_{18} = 4.12 - 3.99 = 0.13.$$

For this university, the student satisfaction score is quite a bit higher than the fitted student satisfaction score estimated from the multiple regression model. (In fact, the student satisfaction score for this university has the largest positive residual in the sample.) The values of student–staff ratio and academic services spend do not allow us to predict the student satisfaction score so well in this case.

Solution to Activity 24

(a) Assumption 2, that the W_is have zero mean and constant variance, can be checked by using a residual plot which plots the observed residuals w_i against the fitted values $\widehat{y}_i$. The residuals should be scattered randomly about zero if the assumption is true.

(b) Assumption 3, that the W_is are normally distributed, can be checked using a normal probability plot for the observed residuals w_i. If the assumption is plausible, then the residuals should lie reasonably close to a straight line.

Solution to Activity 25

With the possible exception of one large positive and one large negative residual, the points in the residual plot appear to be scattered randomly about zero, suggesting that the assumption that the W_is have constant, zero mean and constant variance seems plausible.

The residuals lie reasonably close to a straight line in the normal probability plot, so the assumption that the W_is are normally distributed seems plausible. There is perhaps a hint of curvature, but with only 24 data points it doesn't seem to be sufficient to rule out the assumption of normality.

Solutions to exercises

Solution to Exercise 1

(a) For Forbes's data, S_{xx} and S_{xy} are given by

$$S_{xx} = 10\,820.9966 - \frac{426^2}{17} \simeq 145.938,$$

$$S_{xy} = 86\,735.495 - \frac{426 \times 3450.2}{17} \simeq 277.542.$$

The least squares estimates of β and α are

$$\widehat{\beta} = \frac{277.542}{145.938} \simeq 1.90$$

and

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = \frac{3450.2}{17} - \widehat{\beta} \times \frac{426}{17} \simeq 155.30.$$

The equation of the least squares line is therefore

$$y = 155.30 + 1.90\,x.$$

That is, the fitted model is

$$\text{boiling point} = 155.30 + 1.90 \times \text{atmospheric pressure},$$

where temperature is measured in °F and pressure in inches Hg.

(b) The estimated value of the intercept, $\widehat{\alpha}$, is of little interest in this context because it refers to zero atmospheric pressure, which is of no interest on a mountain and is way beyond the range of the data to which the linear regression model was fitted.

The value $\widehat{\beta} = 1.90$ is the estimated value of the slope. It estimates that for each increase in atmospheric pressure of one inch of mercury, the boiling point of water will, on average, increase by about 1.9 °F.

(c) If the pressure is 25 inches Hg, the predicted boiling point of water is

$$155.30 + 1.90 \times 25 = 202.8\ ^\circ\text{F}.$$

Solution to Exercise 2

(a) On the basis of Figure 28, yes, a linear regression model appears to continue to provide a good model for the full dataset.

(b) This plot shows no particular pattern; the points seem to be randomly scattered around zero. That is, Assumption 2 seems to be satisfied. (Or do you think you perceive a curve to the plot, in which case the linearity of the model would appear to be in doubt?)

(c) The points on the probability plot fall in a pretty good straight line. The assumption of normality does not appear to be in doubt.

(d) Numerically, the slopes of the lines are very similar, but the intercepts are rather different. The lines are plotted in Figure 36.

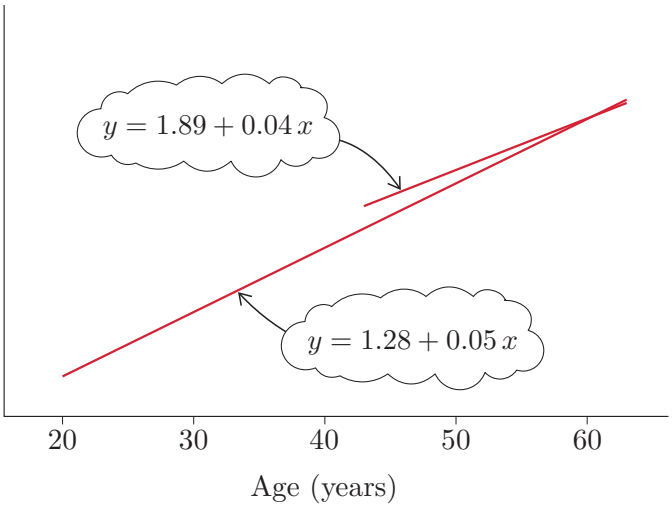


Figure 36 The lines $y = 1.89 + 0.04\,x$ plotted for $43 \leq x \leq 63$ and $y = 1.28 + 0.05\,x$ plotted for $20 \leq x \leq 63$

The fitted lines are similar for the older age range. However, they will clearly differ more for lower ages. So, yes, the line has changed appreciably with the inclusion of younger patients. (In particular, the intercept has changed substantially.)

(Or is there indeed a curve in the residual plot of part (b), suggesting a slightly different, non-linear, relationship over the wider age range? Statistics is full of such ambiguities, especially when arguments are being made, as here, on the basis of small datasets.)

Solution to Exercise 3

(a) First,

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{2.455}{22} \simeq 0.1116.$$

The null hypothesis is $H_0 : \beta = 0$. From Distributional Result (11), the null distribution of the test statistic is $t(n-2) = t(22)$. The observed value of the test statistic is

$$\frac{\widehat{\beta} - 0}{s/\sqrt{S_{xx}}} = \frac{0.05}{\sqrt{0.1116}/\sqrt{4139.77}} \simeq 9.63.$$

From the table in the Handbook, the 0.999-quantile of $t(22)$ is 3.505 so the p-value for a two-sided test is considerably less than 0.002. (In fact, the p-value is very small indeed.) There is therefore very strong evidence against H_0; there does seem to be a relationship between age and cholesterol over the wide range of ages in the dataset.

(b) The point prediction for the value of total cholesterol for a patient with hyperlipoproteinaemia aged 35 years is

$$1.28 + 0.05 \times 35 = 3.03 \text{ mg/ml}.$$

For a 90% prediction interval, we need the 0.95-quantile of the $t(22)$ distribution, which is 1.717. Using Interval (13), a 90% prediction

interval for the value of total cholesterol for a patient with hyperlipoproteinaemia aged 35 years is given by

$$\left(\widehat{\alpha}+\widehat{\beta}\,x_0 \pm t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1}\right)$$

$$=\left(3.03 \pm 1.717\sqrt{0.1116}\sqrt{\frac{(35-39.42)^2}{4139.77}+\frac{1}{24}+1}\right)$$

$$\simeq (3.03 \pm 0.587) \simeq (2.44, 3.62).$$

(c) The prediction interval in part (b) suggests that it is plausible that a 35-year-old individual with hyperlipoproteinaemia would have a total cholesterol level of somewhere between about 2.4 mg/ml and 3.6 mg/ml. Although this is quite a wide range of values, this is useful information since the prediction interval contains only values higher than the observed values associated with some of the younger individuals in the dataset and lower than the observed values associated with many of the older individuals in the dataset.

Solution to Exercise 4

(a) The p-values for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2, 3$, are 0.000, which means that for each of the first three regression coefficients $p < 0.01$. There is therefore strong evidence that β_1, β_2 and β_3 are all non-zero. Also, the p-value for the two-sided test of the null hypothesis $H_0 : \beta_4 = 0$ is $0.032 < 0.05$. There is therefore moderate evidence that β_4 is also non-zero. Therefore there is evidence that the four explanatory variables together influence Y, the rate of growth of GDP.

(b) The points in the residual plot appear to be scattered randomly about zero, suggesting that the assumption that the W_is have constant, zero mean and constant variance seems plausible. Most of the residuals in the normal probability plot lie roughly along a straight line, so the assumption of normality of residuals also seems plausible. Having said that, a number of the larger residuals deviate from the line, so the assumption of normality might be called into question.

(c) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of x_1 increases by one unit, and the values of x_2, x_3 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.0895. (The decrease is because of the negative regression coefficient.)
- Regression coefficient for x_2: If the value of x_2 increases by one unit, and the values of x_1, x_3 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.02118. (The increase is because of the positive regression coefficient.)
- Regression coefficient for x_3: If the value of x_3 increases by one unit, and the values of x_1, x_2 and x_4 remain fixed, then the rate of growth of GDP (y) would be expected to increase by 0.00519. (The increase is because of the positive regression coefficient.)

- Regression coefficient for x_4: If the value of x_4 increases by one unit, and the values of x_1, x_2 and x_3 remain fixed, then the rate of growth of GDP (y) would be expected to decrease by 0.00892. (The decrease is because of the negative regression coefficient.)

Reasons why the regression coefficients for x_1, x_2 and x_3 make sense were given in the solution to Activity 22. The negative regression coefficient for x_4 makes sense because having a high prevalence of HIV can reduce productivity and therefore decrease growth.

(d) Using the fitted multiple regression line, a country with $x_1 = 6$, $x_2 = 25$, $x_3 = 40$ and $x_4 = 0.1$ is predicted to have a growth rate over the ten-year period of

$$0.357 - 0.0895 \times 6 + 0.02118 \times 25 + 0.00519 \times 40 - 0.00892 \times 0.1 \simeq 0.56.$$

Addition of HIV prevalence into the model has not changed the prediction of the growth of GDP of this country (at least, not to second-decimal-place precision).

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 3: © http://ushistoryscene.com/article/rise-of-public-education/

Page 5: © Thanavut Chao-ragam / www.123rf.com

Page 7: © eltpics This file is licensed under the Creative Commons Attribution-Non-commercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Page 8: © Ina van Hateren / www.123rf.com

Page 17: © 2000–2017 vBulletin Solutions Inc

Page 19: © BruceBlaus / https://commons.wikimedia.org/wiki/File:Blausen_0052_Artery_NormalvPartially-BlockedVessel.png This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 21: © Geography Photos / Universal Images Group

Page 25: © 2008 Joyce Gross, University of California, Berkeley

Page 29: © kzenon / www.123rf.com

Page 30: © Sara Riggare

Page 33: © odessa4 / www.123rf.com

Page 44: © pifate / www.123rf.com

Page 46: © Paul Sableman This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 47: © Alan Light This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 51: © edella / iStock Editorial / Getty Images Plus

Page 54: © Zoo New England

Page 56: © kasto / www.123rf.com

Page 58: © kzenon / www.123rf.com

Page 60: © Ilbusca / iStock Unreleased / Getty Images

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.